Exploration, quantification, and mitigation of systematic error in high-throughput approaches to gene-expression profiling: implications for data reproducibility

Robert R. Kitchen

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
to the
University of Edinburgh
2011

Abstract

Technological and methodological advances in the fields of medical and life-sciences have, over the last 25 years, revolutionised the way in which cellular activity is measured at the molecular level. Three such advances have provided a means of accurately and rapidly quantifying mRNA, from the development of quantitative Polymerase Chain Reaction (qPCR), to DNA microarrays, and second-generation RNA-sequencing (RNA-seq). Despite consistent improvements in measurement precision and sample throughput, the data generated continue to be affected by high levels of variability due to the use of biologically distinct experimental subjects, practical restrictions necessitating the use of small sample sizes, and technical noise introduced during frequently complex sample preparation and analysis procedures. A series of experiments were performed during this project to profile sources of technical noise in each of these three techniques, with the aim of using the information to produce more accurate and more reliable results.

The mechanisms for the introduction of confounding noise in these experiments are highly unpredictable. The variance structure of a qPCR experiment, for example, depends on the particular tissue-type and gene under assessment while expression data obtained by microarray can be greatly influenced by the day on which each array was processed and scanned. RNA-seq, on the other hand, produces data that appear very consistent in terms of differences between technical replicates, however there exist large differences when results are compared against those reported by microarray, which require careful interpretation.

It is demonstrated in this thesis that by quantifying some of the major sources of noise in an experiment and utilising compensation mechanisms, either *pre-* or *post-hoc*, researchers are better equipped to perform experiments that are more robust, more accurate, and more consistent.

Thesis Organisation & Structure

The results within this thesis (Chapters 2-5) are presented in the style of international peer-reviewed journal articles. Chapters 2 and 3 have already been published and Chapter 4 has been submitted for review (see below). As a result each of the chapters stand-alone, however the Figure and Table numbers have been modified from the originals such that they fit in the context of the larger thesis. The references have also been modified such that at the end of the thesis there exists a single, unified, reference list. It is acknowledged that an inevitable consequence of this thesis format is some overlap and repetition of background and methodology. However, it is thought the result is a clear and concise thesis summarising the valuable work of this doctoral project.

Chapter 2 R.R. Kitchen, M. Kubista, and A. Tichopad. Statistical aspects of quantitative real-time PCR experiment design. *Methods*. 2010; 50(4):231-6.

Chapter 3 R.R. Kitchen, V.S. Sabine, A.H. Sims, E.J. Macaskill, L. Renshaw, J.S. Thomas, J.I. van Hemert, J.M. Dixon, and J.M.S. Bartlett. Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles. *BMC Genomics*. 2010; 11(1):134.

Chapter 4 R.R. Kitchen, V.S. Sabine, A.A. Simen, J.M. Dixon, J.M.S. Bartlett, and A.H. Sims. Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments. *BMC Genomics*. 2011 (submitted).

Brief details on the collaborators who's work and support was of direct relevance to the material and results contained in this thesis are provided below:

Dr. Jano van Hemert is a supervisor of this project at The University of Edinburgh, UK;

Dr. Andrew Sims is a colleague and mentor at The University of Edinburgh, UK;

Prof. Mikael Kubista and Dr. Ales Tichopad are collaborators at the Institute of Biotechnology AS CR, Czech Republic and the Technical University of Munich, Germany, respectively, and gathered the qPCR data used as the basis for chapter 2;

Profs. John Bartlett and Michael Dixon are collaborators from the Edinburgh Cancer Research Centre, UK, who funded and supervised the clinical study on which microarray analyses were based in chapters 3 and 4;

Drs. Jeremy Thomas, Lorna Renshaw, Jane Macaskill, & Vicky Sabine are collaborators from the Edinburgh Cancer Research Centre, UK, who gathered and processed patients and samples for the microarray analyses presented in chapters 3 and 4;

Prof. Arthur Simen is a colleague and mentor at Yale University, USA, and provided access to the RNA-seq and microarray data used as the basis for chapter 5.

All other practical work and analyses in this thesis is attributed to Robert. R. Kitchen.

Declaration

Except where otherwise stated, the research undertaken in this thesis was the unaided work of the author. Where the work was done in collaboration with others, a significant contribution was made by the author.

R. R. Kitchen
June 2011

Acknowledgements

I would like to thank my supervisors, Jano van Hemert, Andy Sims, Peter Clarke, and Varrie Ogilvie for their kindness, guidance, and tolerance throughout. I've been exceedingly lucky to have such brilliant people involved in my life and my career; I couldn't have wished for better.

I have also had the pleasure to meet (and corrupt) a number of equally excellent people during the last three years and this thesis is a direct result of, and testament to, their wonderful generosity, hospitality, inspiration, and support. I am indebted to Ales, Mikael, Michael, and the postdocs in Munich; Vicky and John here in Edinburgh; and Arthur, Kelly, Libby, Kiran, Marcia, Mark, and Ben in New Haven. Thanks too to Malcolm Atkinson for enabling my summer in the US. I hope that we can continue to work together, and drink beer together, for many years to come.

For keeping me sane and making me laugh for the past three months (and for the years previous), thanks to the students and staff at the MMC, to the Informatics Burgers, and all my friends here in Edinburgh and elsewhere. It's been a pleasure.

It is also no secret that due to Becky's tireless efforts to overcome my ignorance of chromosomes, and a great many other things, she has made me a better person. You're amazing, thanks for everything Miss.

Finally, and most importantly, thanks to my brother and to my parents for their total and ineffable support in anything I've ever chosen to do. This thesis is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General Introduction - Approaches towards quantification of gene expression

1.1.1 DNA, RNA, and gene expression

The human genome is composed of deoxyribonucleic acid (DNA) and contains tens of thousands of individual sequences called genes. The DNA molecule is a double-stranded helix and, at the smallest scale, each strand is an unbroken chain of molecular elements called nucleotides. There exist four types of nucleotide, each containing a phosphate group, a deoxyribose sugar, and one of four nitrogen bases. These bases are adenine, guanine, cytosine, and thymine and are commonly abbreviated to the more familiar A, G, C, and T. The two strands are bound by hydrogen bonds between complimentary pairs of nitrogen bases; with pairing adhering to the rule that A bonds only with T, and C with G. This binding is known formally as the principle of DNA base pairing.

Genetic information is ultimately expressed in the organism through the coding of proteins specific to particular genes. Genetic information stored in the DNA is expressed in two stages: first, the gene is transcribed into the form of messenger ribonucleic acid (mRNA); second, this mRNA is translated into protein. This process of expression is the central dogma of molecular biology in which DNA transcribes mRNA that is translated into protein [1].

An mRNA molecule is unique to a particular gene and is a single-stranded, complementary copy of the nucleotide sequence of that gene, with the slight complication that each thymine (T) base is replaced by uracil (U). Proteins are sequences of twenty different types of amino acids and the translation of nucleotide-triplets, called codons, from a mRNA molecule into these amino acids is specified by the genetic code.

It is the expression of both mRNA and protein that is of interest to biology. Proteins have a much wider range of functions than mRNA, creating molecular complexes that are vital both at the very centre of the nucleus, where they bind to DNA to enhance or suppress transcription, and at the surface of the cell, where they interact with the extracellular environment. Unfortunately, for many genes, the relationship between mRNA and protein expression has been found to be non-linear and this lack of correlation has been the subject of a number of published studies including [2, 3, 4, 5]. Due mainly to technological limitations in protein-level analysis, expression at the RNA-level has instead been a traditional proxy for both levels in analyses of genome-wide gene-expression.

In the remainder of this thesis, only mRNA expression experiments will be discussed. As such, any and all reference to terms such as RNA, expression, and abundance are used interchangeably and in the context of mRNA expression.

1.1.2 Genome-wide expression experiment techniques

> "*Genomics aims to provide biologists with the equivalent of Chemistry's periodic table*" - Eric S. Lander [6].

Several techniques have been developed to allow scientists to detect mRNA transcripts. Early methods, such as Northern blotting [7], required large amounts of RNA and provided, at best, semi-quantitative estimates of expression because results had to be interpreted visually. The past 30 years has seen a rapid development of techniques and instrumentation for mRNA quantification allowing the number of transcripts present in a sample to be measured with increasing accuracy.

Modern research into gene expression relies heavily on mRNA quantification. Current thinking suggests that the expression of any given gene is usually proportional to the number of complimentary mRNA molecules within the cell or tissue. Advances in measurement techniques to allow quantification of transcripts make measuring gene expression in this way an attractive prospect for researchers.

In this thesis, three contemporary techniques by which the abundances of mRNA transcripts can be quantified will be discussed. Introduced in the following three sections, these are quantitative Polymerase Chain Reaction (qPCR), gene-expression microarray, and second-generation RNA-sequencing (RNA-seq). Both microarrays and qPCR measure the amount of mRNA through the annealing of complementary strands of DNA [8, 9], while RNA-seq produces millions of short sequence fragments through a variety of methods that are platform dependent [10, 11].

The major differences between these three techniques are accuracy, specificity, and capacity. Large numbers of genes are concurrently probed by microarrays and as a result the research community has enthusiastically embraced such technologies as a tool for generating data-driven hypotheses. Despite being restricted to analyses of only a small number of genes in a single experiment, the greater accuracy and specificity of qPCR has led to its adoption in smaller, hypothesis-driven, studies including those validating gene-expression levels obtained by array-based experiments. However the throughput of qPCR technologies is steadily increasing, blurring this distinction to microarrays [11, 12]. Due to the higher dynamic range and the option of measuring absolute, as well as relative, abundance of mRNA targets in a sample, if it were not for the high financial cost of qPCR on a per-gene basis, they would threaten microarrays as the researchers' default hypothesis-generating experiment.

RNA-seq provides, for the first time, the capability to directly sequence almost entire cDNA transcriptome(s) contained within a given sample in a high-throughput and affordable manner. It is this ability to estimate expression across the entire length of a target transcript, rather than just within small region targeted by a probe, that

makes RNA-seq such an appealing prospect compared to existing technologies such as microarrays and qPCR [13]. Due to the quantification along entire transcripts RNA-seq has higher specificity than qPCR and number of transcripts concurrently assessed is generally greater than that afforded by arrays, depending on the number of reads obtained from the sequencer.

1.2 Experiment design and bio-statistics

Insight, in the context of gene-expression, into factors governing some biological trait is commonly achieved by obtaining observations of randomly sampled sources/subjects belonging to one or more populations that are representative of the trait under investigation. Expression measurements of one or more genes are summarised in terms of the similarity of all observations between sources/subjects within each population and then assessed for bulk-effects between the populations. Genes that are found to vary little between samples within each population compared to the variation between the populations are thus correlated with the experimental factor(s) and are considered interesting/relevant to the studied trait. There are myriad ways in which the variability between observations of randomly selected subjects within a population can be affected. Such variability is considered as either biological in origin, if it results from some fundamental property of the populations from which samples are obtained, or technical in origin, resulting from a lack of measurement precision in the collection of the observations.

1.2.1 Biological variation

Legitimate variation between individual experimental subjects and samples. The top level of biological variation is dependent on the purpose of the investigation and, at it's most extreme, could represent variability between individuals from a number of distinct species. In most cases, however, biological variation is simply the variability between RNA samples that cannot be explained by the experimental factors under investigation or the technical noise inherent in the experiment.

In this regard, biological variation is entirely a product of the fact that it is not possible to define a set of experimental factors that are capable of accurately predicting the expression of any given gene within any given sample. It is not expected or physically meaningful, of course, that it will ever be possible to make such predictions, but this is only half of the story. Real-world problems arise when the model defining experimental factors is inadequate, including either too few variables or, worse, the wrong variables entirely. In this case the biological variation will be

large and potentially sufficient to obscure the differences between populations defined in the model. Poor experiment design is the cause of inadequate description and implementation of the model designed to describe the differences between populations of individuals or samples.

Of course, poor experiment design does not always equate to bad experiment design, especially in hypothesis-generating experiments. For example, systems that are subject to complex and previously undocumented epigenetic or post-transcriptional expression regulation might exhibit sufficient biological variation to obscure legitimate differences between populations. Indeed, the act of measuring gene expression in living tissue is far from simple, as transcription and translation of genetic information is a highly dynamic and complex process that reflects the biosynthetic needs of a cell's environment, an excellent review is provided in [14].

Therefore, an effective means of reducing the impact of this unmodelled (or unmodellable) complexity is to deliberately obtain a large amount of RNA from a diverse and large set of random samples from each population. This averaging restricts the scope of the analysis to so-called steady-state differences between the chosen populations, which are the least variable between individual cells and subjects, but the most variable between the populations. Unfortunately, the collection of a large set of random and diverse samples is not always possible and in such cases it is usually only possible to reliably resolve large differences between the populations.

1.2.2 Technical variation

Technical variation, on the other hand, is introduced entirely as a result of the, often complex, procedures involved in physically collecting, storing, extracting, and preparing the RNA for analysis. This variability is easily defined, and measured, as that which remains in expressions derived from the repeated measurement of a single biological sample. Exactly as is the case for biological variability, it is not possible to account for all possible sources of technical noise and abstractions must be made to model those sources that account for the majority of the error.

For this abstraction we employ 'surrogate' variables, which may be influenced by numerous measurable or un-measurable factors that are all introduced into the experiment at the same time. For example, atmospheric ozone concentration has previously been shown [15] to affect microarray-based expression measurements. Even though it is possible to record the ozone concentration during an experiment, it is not feasible to do so. It is not the only source of potentially confounding variation that one would have to measure during an experiment, but if it is reasonable to assume that a number of sources, including ozone concentration, ambient temperature, and humidity

remain approximately constant over the course of a day then simply recording the date on which the experiments took place is sufficient to compensate for the bulk variation, caused by the individual contribution of each source, to the experimental measurements.

Another example is the definition of a surrogate for each of the different technicians involved in performing the experiments. A great many factors differentiate people, (including, for example, their skill with a pipette) but are too difficult and too numerous to record individually. Therefore we compensate for the aggregate of all these individual differences between people using the 'person' surrogate.

The difficulty is in the definition of surrogates and how much variation is not catered for in making assumptions such as the consistency of ozone throughout the course of a day. All depends on relative contributions to the total variance by the surrogates that are easiest to measure. The investigation of the relative contribution of various surrogates to the overall technical error is what is of interest in this thesis.

1.3 Reverse-transcription qPCR gene-expression analyses

In quantitative real-time PCR (qPCR) a target double-stranded DNA sequence is exponentially amplified allowing the number of input molecules it to be estimated. Reverse transcription qPCR (RT-qPCR) is an established adaptation of this technique for measuring the quantity of mRNA in biological samples at high sensitivity [16]. Due to this high sensitivity, it is possible to obtain a reliable measurement of genes for which mRNA is in relatively low abundance, even as little as one-cell equivalent [17, 18]. In a typical PCR reaction the cDNA, reverse-transcribed from mRNA in the sample, may be amplified up to forty times, where each round of amplification results in an approximate doubling of product. The 'real-time' aspect of qPCR differentiates the technique from traditional methods that relied on gel-electrophoresis to quantify amplified DNA. Real-time PCR allows monitoring of the amplification during the reaction through the use of fluorescent dyes and is described below.

1.3.1 Targeted amplification of short sequences using primers

Prior to qPCR analysis, all mRNA extracted from a sample is reverse-transcribed to cDNA and entered into the qPCR reaction tube. Primer sequences, designed to complement a specific short region within the target cDNA, define the start and end positions of the short probe/amplicon that will undergo amplification. It is important that the design of these forward and reverse primers is sufficiently specific that only the desired sequence is amplified. The thermodynamic properties of the primer must also be taken into consideration such that hybridisation can occur efficiently during the

reaction.

PCR amplification, first described in [19] occurs in several cycles. In the first cycle, forward and reverse primers are annealed to the positive and negative strand, respectively, of denatured cDNA, before DNA polymerase is used to extend each primer, 5' to 3', producing double stranded DNA starting from the primer location on each strand. In the second cycle the DNA is again denatured and more primers annealed to each of the four single strands, which are again extended using the polymerase. This process of denaturation, primer annealing, and extension continues through subsequent cycles and results in an approximate doubling of the target amplicon after each cycle [19].

However this amplification is not a perfect doubling throughout the entire reaction. At the start of the reaction there are a large number of cDNAs in the reaction mixture that are not targeted by the primers and towards the end of the reaction the primers themselves are depleted. Both of these mean that the reaction curve is, in log-space, sigmoidal, in which the linear 'sweet spot' of almost perfect doubling in the mid-range is where there are plenty of primers and an abundance of target cDNA sequences compared to the other, non-target, molecules in the reaction tube. Different primer sequences have also been found to have different amplification efficiencies, necessitating post-hoc estimation and normalisation between samples/genes during statistical analysis [20, 21].

This amplification is quantified through the use of dual-labeled fluorogenic hybridisation probes [22, 23] in which one of the labels serves as an emitter (a.k.a. 'reporter') and the other as a quencher. When the polymerase binds to and extends the primer, the reporter is released leading to an increase in peak fluorescence at 518nm [18].

1.3.2 Quantification of amplified target-sequences

The most popular method of PCR analysis uses the number of cycles required in each reaction for the fluorescent intensity in the PCR tube to rise above a predefined, and somewhat arbitrary, threshold value, typically based on 10 standard deviations above background in the preceding cycles [18]. This cycle count is referred to as the 'cycle threshold' (Ct) and the smaller this value, the fewer amplification cycles are necessary to attain the critical intensity and the larger the initial quantity of transcript in the sample. Figure 1.1 illustrates the sigmoid fluorescence response curve throughout the PCR amplification and the meaning of the Ct and the threshold.

In general, there exist two main methods for quantifying input mRNA using the Ct data output from RT-qPCR experiments. Absolute quantification uses concurrent RT-qPCR reactions, in different tubes, of known amounts of material at several dilutions

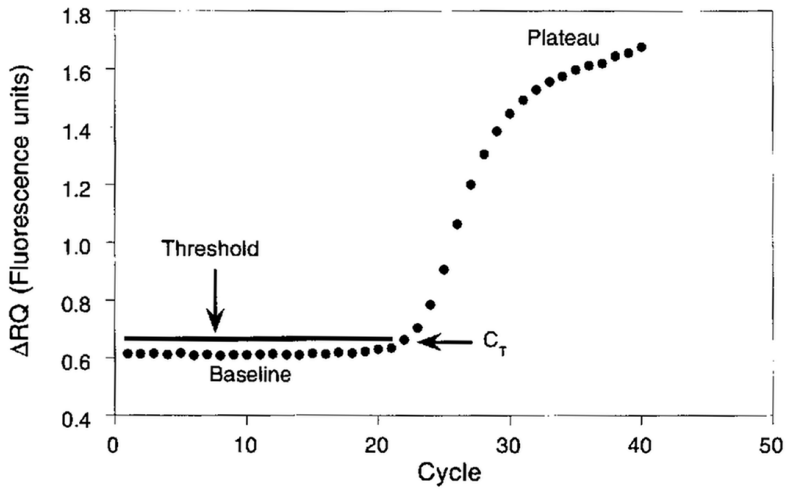


Figure 1.1: An illustration of the qPCR amplification curve, taken from Heid *et al.* (1996) [18]

to calculate a standard curve from which the amount of material in the unknown sample can be extrapolated. Relative quantification uses Ct data from one or more stable reference genes, against which to normalise expression estimates of the gene(s) of interest, to calibrate for total mRNA input and global expression differences between biological replicates [24].

Absolute quantification is more time consuming and costly, in that several dilutions are required for the standard curve and one must have very high confidence in the concentration of target cDNA in the original reference sample, and all its subsequent dilutions for the method to accurately quantify the sample of unknown quantity [24]. In the case of relative quantification, obviously the choice of reference gene(s) with which to relatively quantify the gene of interest is extremely important. An unstable reference gene that co-varies with the treatment groups will lead to a mis-estimation of the difference between groups in the gene of interest [25]. For this reason, a large number of studies continue to be published exploring the most stable battery of reference genes to use for in analyses of a given species or for a given application [26, 27].

1.3.3 Technical variation and efforts towards standardisation

Although technical variation between qPCR replicates is low, variable sample-preparation procedures and conditions, poor experiment design, and poor data analysis have made it difficult to interpret published methods and results from qPCR experiments [28]. Various technical facets of qPCR experiments, such as input cDNA quality, the lab

technician's experience, the efficiency of reverse-transcription and the subjectivity in data analysis have been reported to affect the reliability of qPCR results [25]. To aid in the design, execution, and analysis of such experiments, a standard set of guidelines was recently created that proposes the publication of a minimum amount of information about qPCR experiments (MIQE) [28, 29, 30, 31].

1.4 Microarray gene-expression analyses

Gene expression microarrays allow the analyses of large numbers of transcripts, often thousands at a time, by exploiting the preferential binding of complementary nucleic acids described earlier. This technique is extremely powerful for comparing the expression 'profile' of several samples to detect genes/probes that demonstrate relative changes in expression between sample phenotypes [9], providing insight into gene/phenotype relationships. There are currently a variety of array platforms available with different manufacturers offering different designs and manufacturing techniques, although the basic underlying principal remains constant among them.

1.4.1 Targeted capture of short sequences using probes

Microarrays provide thousands of concurrent gene expression measurements through the quantification of mRNA in each sample using thousands of probes made of single-stranded DNA molecules to which mRNA molecules can attach. These probes can be of variable length, but most modern commercial array technologies involve the use of fixed-length oligonucleotides of 20 to 100 nucleotides [32, 33, 34]. Probes are attached to a solid substrate in a regular pattern, the array, and the location of each probe is recorded for later use [35].

The mRNA in the sample is extracted and labeled with a fluorescent dye before being washed, in solution, over the microarray substrate. Finally, the microarray is washed to remove unbound transcripts and is scanned using a laser causing the dye to fluoresce. Fluorescence intensity is measured using a photo-multiplier tube (PMT) or charge-coupled device (CCD) and a scan over the entire surface of the array produces an image, from which fluorescence intensity values are obtained for each probe [36], usually by means of image analysis software proprietary to the hardware vendor, although open-source alternatives exist [37]. A detailed review of the fundamental considerations of microarray image analyses and information extraction is provided in [38]. The reported intensity values and associated quality information are used as the starting point for further computational analysis.

Probe and array manufacture

Microarray fabrication has improved markedly in the past 20 years, produced by individual research groups to large commercial companies. Early manufacturing procedures to attach probes to the array substrate involved spotting [9], inkjet synthesis [39], and bubblejet synthesis [40]. Systematic variation in the fabrication of the array, for example due to inconsistent amount of probe material deposited over the substrate, was the often the cause of significant error in reported fluorescence values [41, 42].

Affymetrix, a widely popular array vendor, manufacture their 25nt probes through a process of photolithographic synthesis [43]. The significant difference between this and the other methods is that the probes are constructed bottom-up from the substrate, one nucleotide at a time [33]. This overcomes the spatial variability inherent in printing, and a far greater density of oligonucleotide probes can be achieved. However the probes are always built at the same location on every array and are therefore vulnerable to systematic variation in signal intensity over the surface of the array [44].

The probe manufacturing procedure used by Illumina (an array vendor who's technology has been used extensively to generate expression data throughout this thesis), results in the random assembly of glass beads attached to fibre-optic wells in the array substrate [34]. To the surface of each bead is attached around half a million oligonucleotide replicates, comprised of a 50nt probe and a 25nt identifier sequence, illustrated in Figure 1.2. Each bead-type corresponds to a different gene-specific oligonucleotide probe with a unique bead- identifier sequence. Several thousand beads are sampled from an extremely large pool of prefabricated beads and several technical-replicate beads of each type are deposited randomly across the entire surface of the array; with the identifier sequences allowing the eventual location of each probe-type to be determined after deposition [35]. This manufacturing process results in individual arrays that are effectively unique and the random deposition of the beads, and the multiple copies of each bead-type, reduces the impact of spatial variation over the substrate.

Many other microarray vendors exist, as well as bespoke arrays created for specific purposes, however further discussion will be limited to the Illumina and, to some extent, Affymetrix platforms as these are the both widely used and the only arrays used in generation of data for this thesis.

Targeting of probes to the reference

The composition of probes, regardless of their means of manufacture, are the determining factor of biological relevance to the transcriptomic features claimed to be interrogated by the array vendor. Traditionally probes have been designed to

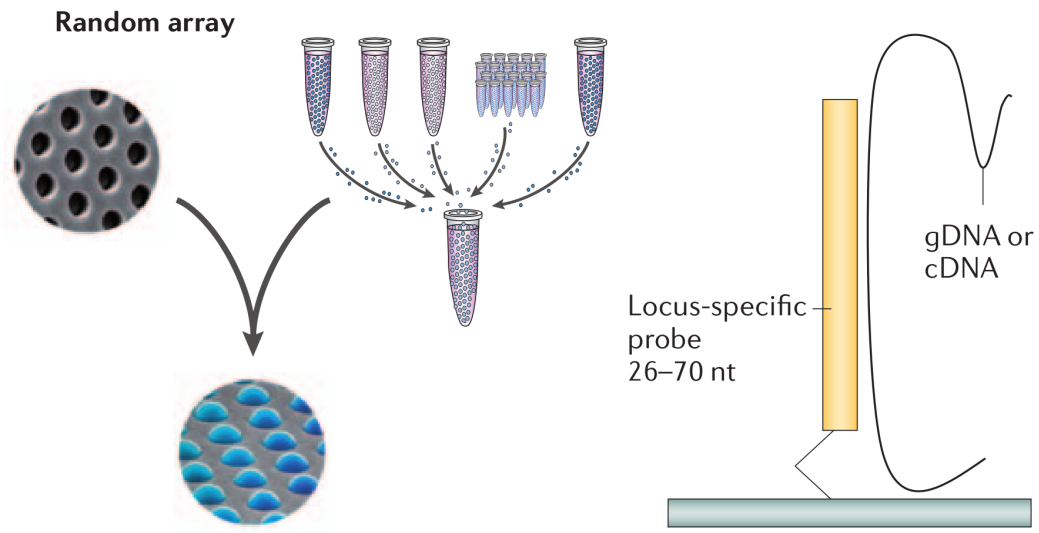


Figure 1.2: Illumina BeadArray manufacturing schematic and illustration of probe hybridisation. Image taken from Fan *et al.* [45]

target regions of genes close to their 3' end, but several other design variants also exist that provide different analysis options. Exon arrays, for example, contain probes targeting every known exon in the transcriptome of an organism [46], others target specific junctions between these exons [47], and so-called tiling arrays contain a large number of probes that target regular intervals over the entire transcriptome [48]. Each has their own specific uses and advantages but for the remainder of this section, and this thesis, only 3' arrays will be discussed.

In general, the longer the sequence of the probe on the array, the greater the specificity to the target region as defined by the reference, however longer probes are more vulnerable to mutations in the sample transcript. Illumina arrays contain long probes and provide multiple observations of each bead-type, but typically contain only around one or two probes per gene. Several negative control bead-types used to compensate for effects such as non-specific binding of cRNA in the sample to random probes on the array [49]. Affymetrix arrays on the other hand, contain many more (shorter) probes and sacrifice multiple replicate copies of each probe-type for multiple observations of each gene and also have deliberately mis-matched sequences. Like the Illumina negative controls, these mis-match probes are designed to assess the specificity of the binding of the probes to the target mRNA. As described in Lipshutz et al. [33], for every 'perfect-match' probe on the array exists a mis-match probe whose sequence

differs from the PM by a single base in the middle of the probe. Typically 11-20 Affymetrix probe pairs are used to interrogating a different location within the target gene and, together, make up what is known as a probeset [50].

Quantification of abundance of captured sequences

Image processing is an important consideration for microarrays and involves calculating foreground intensities using the pixels that make up each feature on the array and estimating a local background intensity using the pixels surrounding each feature. This estimated background is then subtracted from the foreground intensity in preliminary attempt to correct for spatial variation in the brightness of the array [36]. However for features with low foreground intensities, this method can result in negative corrected intensities that must be filtered or adjusted before subsequent analysis. It has also been shown that this background correction increases variability in inter-array measurements and has a detrimental effect on the ability to detect differentially expressed genes [51]. An interesting outcome of the study described in [51] was that there is a correlation between the composition of probes at the nucleotide level and the fluorescent intensity reported by the scanner. Similar observations have also been made in Affymetrix arrays [52]

1.4.2 Analysis considerations

Dual-dye vs. single-dye, sample preparation, and experiment design

Some array designs perform a direct comparison on the chip itself by hybridising a sample from both 'experiment' and 'control' samples, labelled with differently coloured dyes, to the same array. By using two different coloured dyes it is possible to directly compare mRNA levels in two different samples by forcing the RNA from the two samples to compete for the hybridisation to the array probes. Therefore if a gene is more highly expressed in the experiment sample compared to control, the dye used to label the 'experiment' sample will be brighter. The disadvantage of this design is that all information about the absolute expressions of the genes is lost and, as a result, information regarding the individual variation of either the 'experiment' or the 'control' samples is obscured. Furthermore, cDNAs in the sample have different hybridisation efficiency depending on the choice of reporter-dye and a dye-swap is often performed to normalise for this effect. The remainder of the manufacturers produce a single-dye design in which only one sample is hybridised to each array, and the fluorescence of the probes are relative only to the background intensity of the array substrate, therefore preserving the absolute expression levels of each gene under observation. Dual-dye

arrays have special implications for experiment design and the analyses of single-dye arrays is usually simpler [53].

Prior to labelling and array hybridisation, it is important to first assess the quality of the RNA to ensure that results obtained from subsequent expression analyses are due to legitimate biological processes and not to experimental artefact such as, for example, degraded RNA [54, 55]. The protocols for preparing mRNA prior to array scanning is platform dependent, but generally it is the case that mRNA is reverse-transcribed, amplified and labeled with biotin before hybridisation to the arrays.

Filtering, scaling, and normalisation

The removal of 'undetected' probes, which fluoresce only marginally above background, or 'uninteresting' probes, which are not sufficiently variable across all samples to be differentially expressed, is a common and generally beneficial procedure. Such filtering reduces the dimensionality of the dataset, speeding up normalisation and, more importantly, reduces the number of statistical tests performed on the data, lessening the impact of multiple-testing corrections [56, 57, 58].

Despite the utility of filtering, it is very important that the statistic used to determine whether to discard a probe from further analysis does not lead to pre-testing for the downstream differential expression analyses between sample groups [59]. Therefore information used in filtering should not be the same as that used in the downstream testing for differential expression; covered in the following sections on statistical analysis. It is acceptable to use statistics such as the mean expression or the variance of the probe over all arrays, but it is imperative that all are arrays treated equally, making no distinction based on the phenotypes under investigation [59]. Most image analysis software outputs a confidence in the intensity estimate of each feature on the array, or the specificity of the probe hybridisation based on mis-match probes or internal controls; these confidence estimates are a powerful means by which to filter features that are unreliably detected across a number of arrays.

Probes with larger mean intensities have larger variances [60] and data are commonly transformed to reduce this dependence of the variance on the mean. Such transformations can be simply logarithmic and log_2 is a common choice due to the simplicity in interpreting differences between two sets of expressions. The variance stabilising transformation, VST, is similar to the logarithm at high intensities but is a less severe correction at low intensities [61]. VST has been more recently adapted specifically for Illumina BeadArray analysis to make full use of the within-array technical bead-replicates [62].

normalisation is the process of minimising technical variation in measured signal

intensity levels so that biological differences in gene expression can be better resolved. The type of normalisation strategy employed depends upon the expected nature of the technical variation, heavily influenced by the type of arrays used. For example, arrays with printed probes require normalisation for spatial effects, such as that introduced by non-consistent pressure applied by the print head or the physical position of the probes on the array [42]. In the analysis of dual-dye array data with or without a dye-swap, the normalisation might compensate for bulk differences in the signal obtained at each wavelength, for such data a '*loess*' [63] regression normalisation has been recommended, followed by location and scale normalisation [41]; see [64] for a comprehensive review of the treatment of two-colour array normalisation.

In single-channel arrays the normalisation required is simpler as dye intensity or spatial effects due to print head pressure, for example, and are not a concern. Bulk differences in observed intensities between arrays are the major source of obscuring variation for which compensation is required. Such sources include differences in sample preparation (imprecise concentrations, labelling efficiency, etc.), slight differences in the manufacture of the arrays, and the processing of the arrays (different hybridisation time, ambient temperature, scanner differences, etc.) [65]. Several methods have been developed, including fixed-distribution calibration such as quantile [65] (fastest and most widely used); sliding-window based calibration such as loess [42] (slower, less widely used); and methods designed specifically for array expression data such as VSN [61].

Statistical analyses

Much effort has been expended developing algorithms, both open-source and proprietary, to extract valuable results from microarray-derived expression quantifications. Arrays/samples grouped according to phenotype, then tested. A linear model is fitted to each probe-type, within each sample group. Genes are selected based on the significance of the fit of the regression within each sample-group compared to the difference in the sample-group means. Usually this is quantified using a t-statistic, reporting a signal-to-noise ratio of the two distributions of disease and control samples. The outcome of these tests is a list of genes, each exhibiting a statistically significant difference in observed expression between the sample-groups under investigation.

Variations on the basic t-test include *limma* [66], which moderates the variance of each gene towards a common mean calculated using all genes in the dataset using an 'Empirical Bayes' method. This is designed to stabilise variance amongst samples in genes that are very highly or very poorly expressing. Significance analysis of microarrays (*SAM*) [67] was developed to estimate the rate of false positive results

in the output of the standard t-test; which is achieved by permuting samples [68].

1.4.3 Identification of source and scope of confounding variation introduced in array experiments

An early discussion of probable sources of systematic variation included hardware manufacturing issues such as variation in inter- and intra-array quality, lab issues such as RNA sample preparation and hybridisation protocols, as well as array-scanner issues including variation in optical measurements and software image-processing eccentricities [69]. Variation in expression between arrays has been shown to be highly correlated with the concentration of Ozone gas in the localised atmosphere at the time of the experiment [15]. Illumina BeadChips are processed in batches (6, 8, or 12 arrays on a chip), introducing the possibility of systematic batch-effects, non-biological variation in the measured expression undermining the ability to accurately compare samples across different chips and different batches [70]. A comparison of Illumina and Affymetrix platforms found that they yield highly comparable data, although the strength of this relationship is strongest for highly expressing probes and probes that are close to each other in the target gene [71].

A similar investigation attempted to quantify errors introduced at different stages of the sample-preparation workflow on the resulting observations of expression levels [72]. Three levels of variability were identified and corresponded to cell culture, reverse transcription, and hybridisation. The outcome was negative, in that no significant effect was observed as a result of these levels. A comparison of Affymetrix arrays following an RNA amplification protocol designed for small amounts of starting material, to the standard amplification protocol found that direct comparability of expressions was not possible due to the systematic bias introduced by the different protocols [73].

Efforts towards standardisation

A movement towards the standardisation of these levels of data was proposed after it was recognised that combining results across published studies was a highly unreliable process. The 'Minimum Information About a Microarray Experiment', MIAME, [74] aimed to facilitate the interpretation of microarray data and the independent validation of results derived from its analysis. MIAME itself is a standard for recording and reporting microarray data and the authors intend that enforcement of the standard among public repositories of microarray data will enable the development of cross-platform data analysis tools.

The MIAME standard is quite comprehensive; the specification includes detail about the reporting of important elements at all stages of a microarray experiment,

from sample preparation to the method and variables employed in the detection of differential expression. However the standard is interpreted differently by various repositories and journals, and does not mandate some crucial information that might be of use to diagnose systematic batch effects in the submitted data such as date of sample-preparation, hybridisation, and scan.

1.5 RNA-seq gene-expression analyses

The completion of the International Human Genome Project in 2004 [75] has had an enormous impact on the field of biomedical research and has recently been the subject of an excellent review [76]. The enormous investment of time and money in this project were a result of the limitations of Sanger-based capillary sequencing [77, 78] and rapid technological innovation in the years since allow much more ambitious sequencing projects to be performed in a fraction of the time and at a fraction of the cost [79].

During this period of innovation, several competing methods were pioneered and have been extensively reviewed [80, 81, 11, 82], but all essentially share a common underlying approach in which many millions of short sequence fragments are sequenced in parallel. This so-called 'massively parallel', 'next-generation', or 'second-generation' sequencing is achieved through the ligation of short adapter sequences to each end of the fragmented input DNAs which are used to secure the sequences to a solid substrate prior to amplification and sequencing [83].

As was the case with arrays previously, second-generation sequencing methods have been adapted to serve a variety of applications including genome-wide association studies [84], chromatin analysis [85], DNA methylation analysis [86], and RNA analysis [87]. The integration and analysis of these various sources of data is a fantastic challenge and a major area of interest, reviewed in [88]. However, of direct interest in this thesis is the application of second-generation sequencing to the study of gene expression and the quantification of mRNA.

Sequence-based analysis of mRNA has been a viable technique since the advent of Sanger sequencing through the sequencing of cDNA or EST libraries [89]. However, for all but the most ambitious initiatives the sequencing of entire transcripts was prohibitive and, as a result, methods of sequencing a reduced set of much shorter mRNAs were more widely used [45]. These methods included Serial Analysis of Gene Expression (SAGE) method [90] and Massively Parallel Signature Sequencing (MPSS) [91], in which short (10-22nt) tags are obtained from predefined locations in each mRNA molecule and subsequently sequenced. Several studies have compared expression estimates and results obtained from these sequencing methods to those gathered by hybridisation-based microarray technologies, reporting moderate correlation in absolute expression

[92] but that within-technology expression ratios between sample groups are more consistent [93].

Sequence-tag-based and hybridisation-based approaches to gene-expression analysis suffer from a fundamental reliance on an accurate and well-annotated reference with which to make sense of the results. As discussed previously in this chapter, array probes are designed to deliberately target specific mRNAs based on their reference sequence and while MPSS and SAGE methods do not require the a-priori definition of target sequences, the short tags output still require mapping back to the reference before their abundances can be estimated. They also suffer similar biases due to probe/tag position within the transcript [94] and are both liable to mis-interpretation due to alternative-splicing [95].

Second-generation RNA-sequencing (RNA-seq) allows, for the first time, the comprehensive full-transcript, high-specificity, sequencing previously afforded by the Sanger method but with the high-throughput and low-cost of tag- and hybridisation-based methods [87, 96].

1.5.1 Non-targeted fragmentation and sequencing of whole transcripts

To be compatible with most second-generation sequencer technologies, typically large mRNA molecules are selected, reverse transcribed, and randomly fragmented such that the length of the cDNAs are <500nt [97]. These fragments are typically amplified by PCR depending on input amount, however this has been reported to be a potentially significant source of bias, especially in GC-rich genomes [98]. Certain methods of reverse-transcription have also been shown to exhibit detectable bias in the set of reads output from the sequencer [99], and methods of directly sequencing unamplified RNA have recently been proposed to address these biases [100]. Furthermore, the use of RNA, or single-stranded cDNA, are among several methods used to preserve strand-information from the sequencing [101, 102]. Whatever method of library preparation is used, platform-specific adapter sequences are ligated to the input fragments, which are subsequently and immobilisation on a solid substrate. Finally, the fragments are amplified on the substrate in clonal clusters such that a reliable signal can be detected.

The most basic output from second-generation sequencers are images corresponding to each 'cycle' of the sequencer in reading a predefined number (30-200) of nucleotides, concurrently, from the start of each cDNA fragment. For example in the Illumina sequencing method, these cycles relate to the incorporation of fluorescently labelled nucleotides that are chemically blocked such that only one nucleotide incorporation event occurs per fragment population per sequencing cycle [10]. These images are

analysed and the fluorescence signal for each base is quantified along with an estimated confidence based on the detection of the signal above the local background. These quantifications are used for base calling and the more meaningful output of the cDNA 'read' sequences is returned along with the detection quality for each nucleotide. The challenge then exists in mapping these short reads, themselves a small subsequence of the fragmented cDNA, back to the genome of transcriptome of the relevant organism before counting mapped reads to estimate the abundance of the relevant transcript in the original sample. For a schematic overview of these stages of cDNA preparation, sequencing, and mapping see Figure 1.3.

Quantification of sequence-fragments ('reads')

There are a great many methods for aligning these short reads back to the organism reference, which have been discussed in great detail [103, 104, 88], as well as emerging methods of reference-agnostic de-novo assembly [105]. The advantage of RNA-seq compared to hybridisation-based expression analysis methods is that the detection quality of the output sequences is available at a per-nucleotide level, allowing for bulk exclusion of reads failing a certain minimum threshold and advanced alignment with weights allowing small insertions or deletions, 'indels', in regions where the read-quality is reportedly poor [106]. Such tolerance of indels have been extended to the point where full gapped alignments can be performed to identify reads spanning both known and novel exon-junctions, aiding discovery of novel isoforms [107, 103, 108]. Accurate, fast alignment of large numbers of short reads to the reference presents not only a considerable computational challenge but, due to the complexity of the transcriptome, can also have a dramatic effect on estimated expression levels.

Once reads have been assembled or mapped to the reference, the abundance of RNA in the input sample corresponding to the 'region of interest', be it a gene, transcript, or exon is approximately equivalent to the read-depth. This read-depth is the average, over all bases in the given reference region, number of reads in which each base is present. Depending on the choice of RNA/cDNA preparation, the distribution of reads over the length of a gene, transcript, or even an exon is not consistent and exhibits strong bias against the extremes of the region [87]. This presents further challenges to the identification, assembly, and quantification of novel isoforms as most rely on the assumption of consistent expression across the entire length of the transcript. An illustration of the process of translating millions of short reads to expression estimates and insight is provided in Figure 1.4.

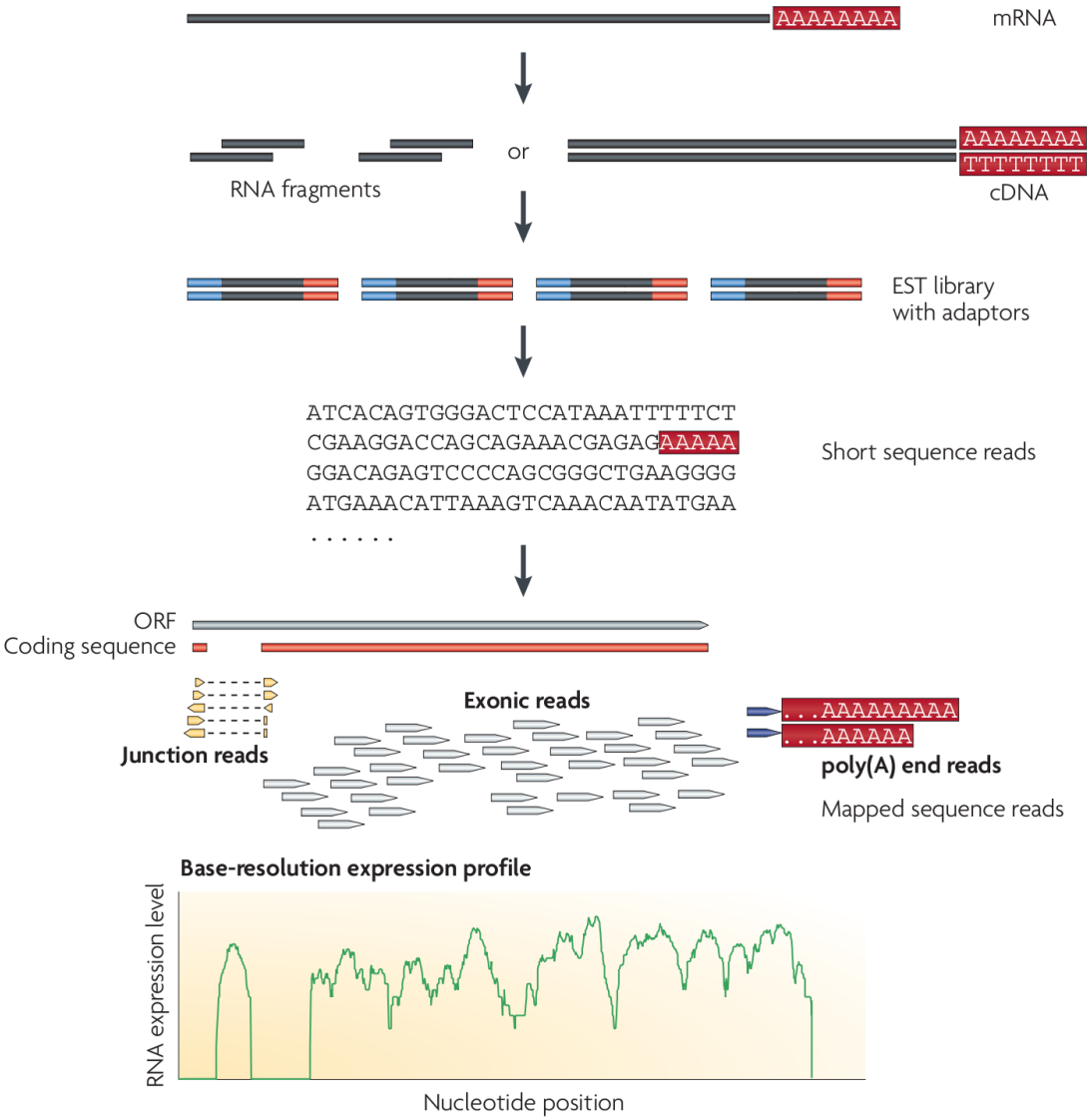


Figure 1.3: Illustration of an RNA-seq experiment workflow; illustration taken from Wang *et al.* [87]

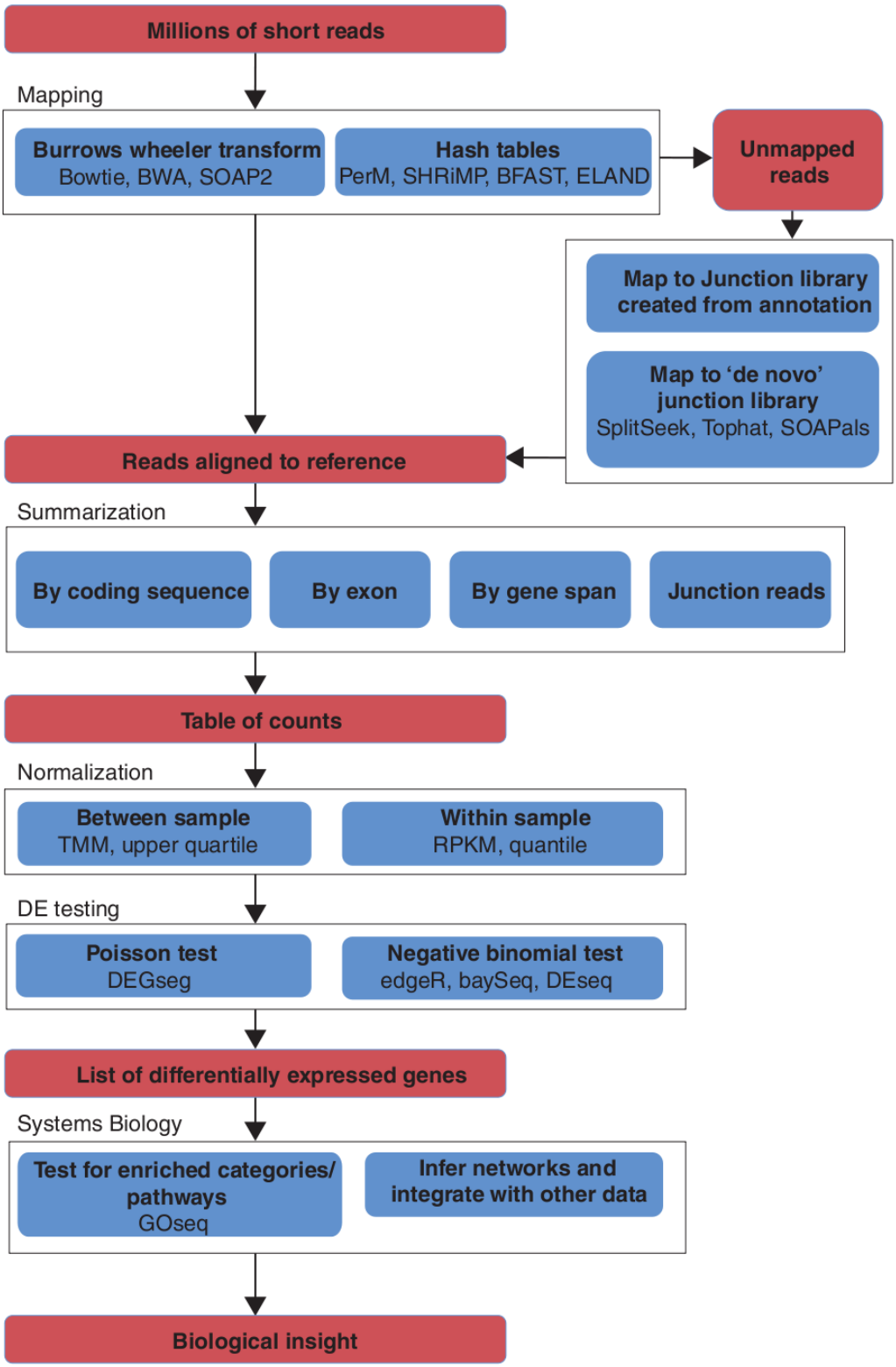


Figure 1.4: Schematic RNA-seq analysis pipeline from Oshlack *et al.* [13]

1.5.2 Analysis considerations

Quantification of this abundance is achieved simply through the counting of the reads mapping to a given region/feature of the genome. Despite the inherent biases in the technique the number of reads mapping to a region of the reference between technical replicate sequencing-runs is assumed to be, and has been shown to be, Poisson distributed [109, 110, 111].

Filtering and normalisation

One of the main advantages of RNA-seq over microarrays is the ability to detect genes and exons of much lower abundance in the input sample. This lower detection limit depends on the number of reads obtained from a sequencing run and the size of the transcriptome of the organism under study. Based on the assumption of Poisson noise between technical replicates, and therefore the minimum possible variability between biological replicates, it is possible to define a filter to remove features for which the observed number of reads in all sample-groups cannot result in a significant test-statistic in analyses of differential expression. Such a filter can be achieved, in a two-group test for differential analysis, by inputting the two total library sizes for all samples in each group into a contingency table prior to a Fisher's exact-test or a Chi-square test. The remaining two values in the 2x2 contingency table are set to zero and iteratively increased for one of the groups only until the test is significant to some predefined value. The number of reads for which the test is significant is then the minimum number of reads across all samples, independent of sample-group and any feature with fewer reads is removed from further analysis.

The total number of reads output by RNA-seq between lanes and samples can be highly variable. Therefore count data are routinely normalised by the total number of reads obtained per sample, due to the obvious situation in which a sample with twice the number of mapped reads is likely to have twice the number of reads mapped to any given feature; in this situation the feature is not biologically differentially expressed, as nave count differences would suggest.

Count data can also be normalised by feature length to account for the fact that longer features will have mapped more reads, so read counts are divided by the number of bases in each mRNA region as well as the total number of mapped reads; first introduced in [107] as 'Reads Per Kilobase of exon-model per Million mapped reads' (RPKM). However this correction has been reported to introduce error in the variance such that the problem of feature-length bias persists [112], their solution being to assess only fixed length regions for expression analyses however this presents obvious issues surrounding the selection of such regions as well as neglecting a potentially huge amount

of expensive experiment data [110]. Also, the choice to perform analyses at the exon-, rather than the gene-level do not overcome length-bias issues as the variation in feature length is approximately the same at both levels [112]. Normalisation has been identified as having a large impact on the detection of differentially expressed features, especially in situations involving a small number of very highly-expressed features [110].

Statistics for resolving differential expression and sources of technical variation

The first volley of articles publishing results of differential expression analyses from RNA-seq count data assumed Poisson variance not only between technical replicates but also between biological replicates [109, 113]. However the variance estimated by such a model has been found to under-represent the variability between biological replicates [114, 115]. Perhaps a better model for these replicates, that incorporates the over-dispersion of the count data, is the negative binomial and this has been adopted by a number of more recent analysis methods [116, 117]. However, determining the correct model of variance in RNA-seq analysis is still an area of active research.

Given that RNA-seq is a nascent technology and that experiments are expensive, there have been limited numbers of studies investigating technical aspects of the reliability and reproducibility of the expressions obtained using this technique. However, technical replicate samples have been assessed in a handful of studies, each finding that the variability between multiple sequencing runs of the same library is very low [109, 110, 111]. Bullard *et al.* noted in their analysis of fold-changes between different tissues in the MAQC stage 3 RNA-seq dataset [118], that estimated variability introduced in replicate library preparations is much smaller, on average, than biological differential expression [110].

1.6 Batch-effects, bias, and systematic error

High-throughput experimental techniques in biology, such as those briefly introduced in the previous sections commonly require a complex set of procedures to prepare biological material for analysis. These procedures vary depending on the platform, the manufacturer, and the current state-of-the-art approach based on the scientific literature. Each stage of sample preparation increases the likelihood that conditions vary during the course of the experiment, introducing systematic variation into the experiment output. Such variation can be both biological and technical in nature and effectively introduces new hidden variables to the model that explains the observed data in addition to those variables that are known, corresponding to different populations

of experimental subjects [119].

The consequences of such batch effects are that the 'new variables', which are unrelated to the biological variables under study, serve to confound interpretation of the experimental results and, depending on the design of the study, can lead to incorrect conclusions being made pertaining to the biological variables. Such a situation is obviously exacerbated if the hidden batch variables are completely confounded with the biological variables. For example, if all observations of a condition are made in such a way that the biological replicates within each condition are contained within different batches. In this hypothetical example, if all samples from treatment group A were processed in January and all samples from B processed in March then if there were some hidden batch effect resulting from the time difference between the observations, then it would be very difficult to differentiate the batch-variation from any interesting biological variation between the treatments.

Near-complete confounding can occur in meta-analyses of a merged set of multiple independent datasets where samples are likely to have been quantified at different times, using different protocols, and on different platforms. Meta-analyses are often valuable means of extracting new information from existing data, and are performed in the hope of attaining increased statistical resolution for detecting differential expression, provided by a large number of samples compiled from several complementary studies. Such analyses often avoid issues surrounding interpretation of independent analyses of differential expression over all constituent datasets, performing instead a single analysis for differential expression following data integration at the expression-level [120]. Effective strategies for normalising data between these sub-experiments are vital, however, to minimise these batch effects and maximise the resolution of the analysis to the effect of interest.

Individual studies, as well as meta-analyses, can be vulnerable to batch effects and depend on the quality of the design of the experiment. Specific examples of published studies have been reported in which batch effects have been found to be completely confounded with the biological contrast [121, 122]. A re-analysis of a microarray dataset published by Spielman *et al.* [122], revealed that the two human populations that were the subject of a gene-expression comparison were completely confounded by processing date and that the large number of genes originally found differentially expressed could not be resolved after compensation for the processing batch [123].

1.6.1 Methods for the detection of systematic error

The task of detecting systematic error is made simpler with prior knowledge of the likely sources of technical variation in an experiment. Analyses using technical

replicate samples are important in providing a basis for assessing the relative impact of confounding variation introduced during each of the stages that lie between the subjects under study and the acquisition of quantified data. Such information can be used as the basis for recommendations pertaining to the optimal design of the study such that technical variation is minimised.

Batch effects not restricted to, but are easiest to identify and model in high-dimensional datasets. Detection of outlying samples and large batch effects can often be performed visually through assessment of the per-sample expression distributions or their similarity via unsupervised dimensionality reduction such as principal components analysis (PCA) [124, 125]. A recent article published by Leek *et al.* [119] attempt to formalise the analysis workflow for diagnosing potential systematic variation using diagnostic methods such as these.

1.6.2 Pre-hoc defense against batch effects

In self-contained experiments, sound experiment design is the most effective strategy to protect against the introduction of confounding technical noise and batch effects. There remains, therefore, the need to have a thorough understanding of the experimental method in order to design it in such a way that the major sources of technical variation are averaged-out in the downstream data analysis.

Such a design would benefit, wherever possible, from 'blocking' samples in a balanced design such that there is the same number of samples from each phenotype-group in each potential batch. For example, in an analysis of samples belonging to two treatment groups using a given experimental technique, if there was good evidence for the introduction of technical variation due to the different days on which the experiments are run then a prudent blocking scheme would be to simply run equal numbers of samples from each treatment group on each day. Clearly, more complicated contrasts require more careful blocking and such experiments will always be vulnerable to introduction of technical noise from unexpected sources.

In addition to unexpected sources of technical variation, there are experiments in which deliberate blocking is not possible. For example, the analyst obviously has no control over the design of existing datasets to be combined in a meta-analysis, nor is it always possible to perform such blocking in clinical diagnostics, in which a training set was generated long before individual samples being classified. In such situations it is often necessary to identify and correct for sources of confounding technical variation after the data have been generated.

1.6.3 Post-hoc compensation for batch effects

Systematic variation often violates assumptions made by standard bulk-normalisation methods. For example, normalising a set of microarray samples such that they all share a common expression distribution assumes that all probes are affected equally by whatever systematic effect caused the distributions to differ in the first instance. This is not always the case, however, and the signal from different genes or probes are frequently affected differently [119]. In such circumstances more invasive, gene-wise, corrections are often required to remove confounding technical variation.

A number of computational methods have been developed for compensating for batch-effects in both large and small datasets. For example, mean-centring batches, for each gene, with or without within-batch variance normalisation, is the simplest and fastest means of correcting for batch-effects [126]. The number of samples in each batch should be fairly large to obtain reliable estimates of within-batch mean and variance; also, care must obviously be taken to account for the biological sample-groups during the correction and the use of within-batch variance scaling is not valid in unbalanced designs [127]. *ComBat* computes an 'Empirical Bayes' shrinkage of the individual gene-wise batch effect parameters, using all genes within each batch, prior to adjusting each gene to compensate for the batch effects [127]. The use of this common shrinkage allows more robust parameter estimation even when the number of samples per batch is small, as is also the case in methods for estimating differential expression in microarray [66] and RNA-seq analyses [117]. Surrogate Variable Analysis (SVA) [128] is a method for estimating technical variation where the source is unknown or known sources do not account for the majority of the experiment noise [119]. Other methods also exist, such as singular-value decomposition [129] and distance weighted discrimination [130], although these methods require a large number of samples per batch in order to obtain reliable results [127].

1.7 Project aims

Before any advanced analysis or biological interpretation can be made from gene-expression experiments, one must have confidence that the observations themselves are faithful to the underlying biology and not a product of confounding variation caused by the specific circumstances in which they were obtained. In many clinical and experimental studies, replicate samples are often unavailable due to either cost constraints, a lack of sufficient experimental subjects, or small starting amounts of material [131]. It is therefore desirable that experiments investigating gene expression subject to these restrictions are as well designed as possible. This requires

reliable knowledge of potential sources of technical variation to better resolve true biological variation in the data. Also, for meta-analyses it is critical to understand the various factors influencing individual datasets being merged, how different the platforms/sample preparation procedures can be before the technical differences between individual datasets starts to negatively affect the power of the meta-analysis.

The previous sections have introduced three of the most common high-throughput methods of assessing mRNA abundance and provided some insight into the means by which these abundances are obtained and the various sample preparation requirements for each. The analyses in the following chapters will focus on ascertaining, for each technology, the reliability of such measurements based on evidence obtained from bespoke experiments and publicly available datasets.

1.7.1 Specific aims

1. To identify and quantify sources of technical variation in three popular methods for assessing mRNA expression; qPCR, microarray, and RNA-seq;

2. Explore the extent to which such technical variation affects results of differential expression analyses;

3. To investigate methods of correcting for observed variation, either through improved experiment design, better data quality control, or statistical corrections.

Chapter 2

Statistical aspects of quantitative real-time PCR experiment design

Robert R. Kitchen, Mikael Kubista, Ales Tichopad.
Methods. 2010 Apr; 50(4):231-6.

Preface

The content of this chapter is exactly that presented in Kitchen *et al. Methods* 2010. The presentation has been modified so as to conform to the formatting guidelines for a Ph.D. thesis chapter. The original article, formatted as it appeared in *Methods*, can be downloaded free of charge from the publisher's website.

Both the original article and this chapter were drafted by myself and, bar a few recommendations by my fellow authors and reviewers, the structure and content is my own. As is covered in the text, the analysis method was previously introduced in Tichopad et al. 2009 where the nested-Anova was applied to the stages of qPCR experiment design jointly by myself and Ales Tichopad. The software introduced in that article, and more fully described in this chapter, was entirely created entirely by myself to implement the nested-Anova. The data presented in the Methods 2010 article and in this chapter, concerning the effect of normalisation to a reference gene on the measured variance components, was analysed by myself.

Abstract

Experiments using quantitative real-time PCR to test hypotheses are limited by technical and biological variability; we seek to minimise sources of confounding variability through optimum use of biological and technical replicates. The quality of an experiment design is commonly assessed by calculating its prospective power. Such calculations rely on knowledge of the expected variances of the measurements of each group of samples and the magnitude of the treatment effect; the estimation of which is often uninformed and unreliable. Here we introduce a method that exploits a small pilot study to estimate the biological and technical variances in order to improve the design of a subsequent large experiment. We measure the variance contributions at several levels of the experiment design and provide a means of using this information to predict both the total variance and the prospective power of the assay. A validation of the method is provided through a variance analysis of representative genes in several bovine tissue-types. We also discuss the effect of normalisation to a reference gene in terms of the measured variance components of the gene of interest. Finally, we describe a software implementation of these methods, *powerNest*, that gives the user the opportunity to input data from a pilot study and interactively modify the design of the assay. The software automatically calculates expected variances, statistical power, and optimal design of the larger experiment. *powerNest* enables the researcher to minimise the total confounding variance and maximise prospective power for a specified maximum cost for the large study.

2.1 Introduction

2.1.1 The importance of experiment design

The typical quantitative real-time PCR (qPCR) experiment is designed to test the hypothesis that there is no difference in the expression of a gene between two or more subpopulations; this is based on experiments performed on representative groups of biological subjects that, for example, exhibit different phenotypic traits or have been exposed to different treatments [16, 132, 133, 134, 135]. If this hypothesis is unlikely to be true, the alternative hypothesis is supported stating that there is a difference between the subpopulations. The ability of the researcher to obtain a statistically significant result from the testing of these hypotheses is governed by three factors:

1. the treatment effect, that is the magnitude of the mean differential expression between the chosen subpopulations;

2. the inherent and expected biological variability in the expression of the gene between subjects randomly selected from within each subpopulation;

3. the technical noise introduced through measurement error.

The larger the treatment effect, the easier it becomes to resolve from the confounding noise. Biological variability is generally unavoidable, but one can seek to minimise its impact by randomly selecting large numbers of subjects (biological replicates) from each subpopulation. Technical noise can be minimised through careful lab practice, the use of technical replicates, and the addition of appropriate controls [136].

The concept of treatment effect and measurement variability is the basis of statistical power. The power of a statistical test is the probability of rejecting the null hypothesis, given that the null is false and the alternative hypothesis is true [137]. In other words, the power is the biological resolution of the experiment; it quantifies the likelihood of being able to resolve any differential expression between treatment groups based on the variance of available measurements. Power increases with increasing magnitude of the differential expression, increasing number of biological replicates, increasing measurement precision, and decreasing biological variability. The objective is to maximise the statistical resolution of the assay, by minimising the confounding variance in the measured experiment data, such that a determination of the treatment effect can be more confidently reported.

It may sometimes be the case that results collected with an assay appear more reproducible where small numbers of biological and technical replicates are employed. This apparent increase in precision and power is illusory, however, and significant results may simply reflect the chance fluctuations in the particular subjects or measurement

processes chosen for the experiment [53]. It is generally considered good experimental practice to vary the conditions of the assay, by sampling multiple subjects and analysing multiple technical replicates, to increase the chance that the statistical significance of the results obtained is real and reproducible in different settings [138].

2.1.2 qPCR experiment design and error propagation

Between the selection of subjects from the subpopulations and the gathering of expression data by qPCR there are several steps of sample-preprocessing that are necessary to prepare the genetic material for analysis. These procedures, illustrated in Figure 2.1, are typically:

1. the sampling of material from each subject and the extraction of the nucleic acid;

2. in the case of RNA analysis, the reverse-transcription (RT) of RNA to convert it into cDNA;

3. the amplification of the cDNA by qPCR.

Some protocols may include additional steps such as fixation of the sample, transportation, and storage. All of these procedures are susceptible to the introduction of error [25] and, combined, they represent the technical noise in the obtained RT qPCR measurement. In addition to the biological variability between subjects, these sources of technical noise all contribute to the total variance of the measured expressions reported by the qPCR. The minimisation of this variance can be achieved through effective, informed experiment designs and sampling plans that employ replicates where they are expected to have the greatest benefit. The challenge is therefore to design experiments with the optimal number of biological and technical replicates such that the statistical power is maximised and sufficient to test a biological hypothesis, while maintaining an affordable and realistic sampling plan.

It is assumed that technical noise introduced into the experiment from each subsequent stage in the sample-preprocessing procedure is independent and, as such, the effect on the overall noise of the assay is additive. Namely, the magnitude of error introduced due to pipetting, uncertainties in instrument readings, and chemical noise in the different processing steps are not considered to be interdependent. There are, however, a few exceptions where this assumption is invalid; for example interference due to the presence of an inhibitor may not be independent if the same inhibitor impairs the performance of several steps of the sample processing.

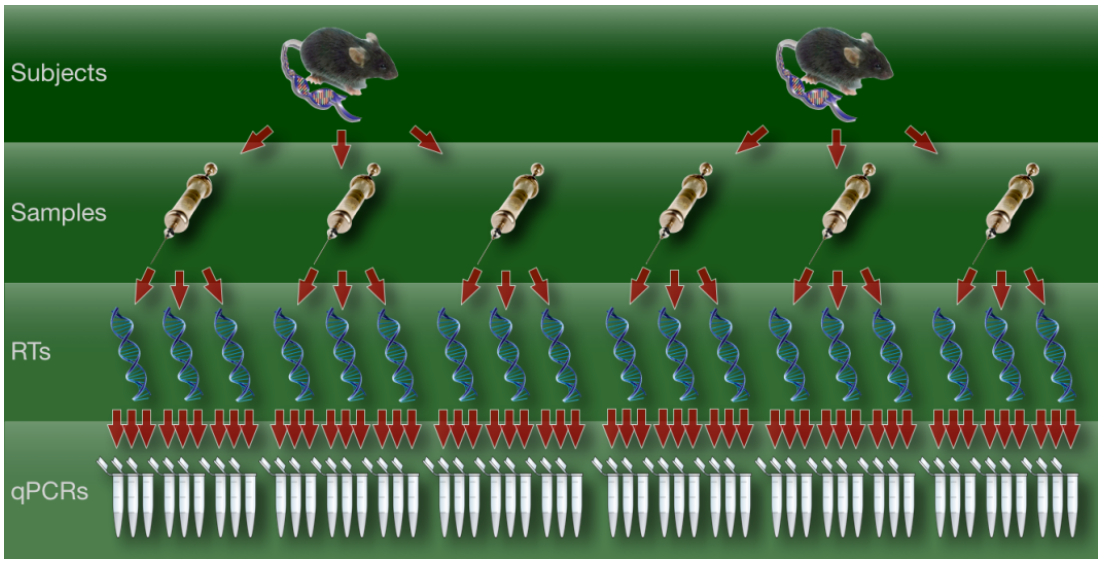


Figure 2.1: Example 2x3x3x3 experiment design for 2 subjects belonging to a single subpopulation. 3 qPCRs are performed for each of 3 RTs of 3 samples from each subject. The result is 27 Cq measurements for each subject, 54 in total for the subpopulation. From this design, variance components at each stage of sampling can be estimated. *Mouse image appears courtesy of the Wellcome Trust Sanger Institute.*

2.1.3 Focus of the paper

Due to the large scope for the introduction of error into qPCR measurements and results, it is not only essential that experiments are well conducted and validated, but also that they are carefully designed and documented; this enables the researcher to maximise the likelihood of accurately and reproducibly reporting interesting biological phenomena [28]. Power analyses afford the researcher a valuable tool with which to estimate the resolution of the assay, in terms of its ability to test a specified hypothesis while the experiment is still in the design phase. The importance and utility of these prospective power analyses is universally accepted, however the calculation of power is often reduced to mere guesswork due to the fact that, by definition, the magnitude of the effect to be studied and the measurement variation in the prospective assay cannot be precisely known [139]. After the assay has been performed and data are available, however, these variables are known (at least in terms of the set of samples analysed) allowing a more accurate calculation of power. Such retrospective power calculations have been shown to be useful, although the measured effect size is often less informative than the variance estimated from the data [140].

In this paper we describe a method of estimating the components of biological variation and technical noise directly from qPCR measurements. This is achieved

through the exploitation of a small number of biological and technical replicates at each stage of the sample processing procedure; these stages being the inter-subject, inter-sample, inter-RT, and inter-qPCR. These biological replicates form a pilot study to a larger, prospective investigation and, as such, should be drawn randomly from a larger cohort of subjects that are to be used as the basis of the future investigation. The variance components estimated from this small pilot study are used to determine the optimal experiment design and sampling plan for the subsequent prospective study. We further exploit these measured variances by including them in the prospective power calculation for the future study, providing a more accurate, evidence-based, estimate of the expected experiment error.

As a validation of the method, variance components are estimated for several genes from a number of bovine tissue-types and contrasted with the components from the same data following normalisation to a reference gene. We use these data to qualitatively assess the utility of technical replication at only the qPCR level; a common practice in qPCR assay design and one for which the rationale is unclear, except perhaps as a low-cost insurance against a failed PCR.

Finally, we present *powerNest*; a software implementation of these methods that provides an intuitive and efficient means of optimising the sampling plan given the data from such a pilot study.

2.2 Description of method

2.2.1 Model

We define the model for any given Cq measurement by qPCR based on four processing steps that account for both the biological variability and the technical noise that influence the measured value. These are the choice of subjects from the subpopulation, the replicate samples extracted from each subject, the replicate RTs of mRNA from the same sample, and the replicate qPCRs of cDNAs from the same RT tube. These effects are all assumed to be independent and randomly drawn from normal distributions on the logarithmic scale [141, 142].

Although the introduction of biological and technical noise at each of the sampling levels is independent, the observed variances are not. The variation introduced at a given level propagates additively throughout subsequent levels, allowing these variance contributions to be modelled. All factors therefore meet in unique combinations and so a nested, or hierarchical, model of additive noise is applied to the measured Cq values [143] such that any given measured Cq can be expressed as

$$Cq_{gijkl} = \mu_g + \alpha_{gi} + \beta_{gij} + \gamma_{gijk} + \delta_{gijkl}, \tag{2.1}$$

where μ_g is the geometric-mean expression of the gene in the g^{th} subpopulation (which is equivalent to the arithmetic average of the fold-change or expression of the gene on the log-scale), α_{gi} is the random effect of the i^{th} subject in the g^{th} subpopulation, β_{gij} is the random effect of the j^{th} sample extracted from the i^{th} subject in the g^{th} subpopulation, γ_{gijk} is the random effect of the k^{th} RT reaction from the j^{th} sample extracted from the i^{th} subject in the g^{th} subpopulation, and δ_{gijkl} is the random effect of the l^{th} qPCR from the k^{th} RT of the j^{th} sample extracted from the i^{th} subject in the g^{th} subpopulation.

This model was previously justified in terms of its application to qPCR experiment design in [144]. The total variance of any given Cq measurement follows directly from the model in Eq.(2.1) and is defined as

$$\sigma^2_{Cq} = \sigma^2_G + \sigma^2_I + \sigma^2_J + \sigma^2_K + \sigma^2_L. \tag{2.2}$$

In simpler terms, the expected variance of each measurement can be divided into two categories; the first is the treatment variation between subpopulations that is expressed by the σ^2_G term; the second is confounding biological variance and processing noise that is encompassed by the sum of the remaining variance components corresponding to inter-subject, inter-sample, inter-RT, and inter-qPCR variation. To maximise the statistical power of the assay one should minimise the confounding variance to be able to accurately resolve the treatment effect.

The variance model in Eq.(2.2) is used to define a nested analysis of variance (nested-ANOVA) that produces estimates of each of the four modelled components of variance. The calculation of these variance components is performed as described in [145], by a procedure essentially based on the subtraction of the sum-squared variations of each level from that of the respective immediate higher level.

The relative contribution of each component, vc_x, to the total variance is expressed as a percentage:

$$vc_x = 100\frac{\sigma^2_x}{(\sigma^2_I + \sigma^2_J + \sigma^2_K + \sigma^2_L)}, \tag{2.3}$$

where $x = I$, J, K, or L.

2.2.2 Experiment optimisation

In terms of the optimisation of the experimental design, it is the objective to minimise the total expected technical and biological variation within each treatment group, g,

which is defined as

$$\hat{\sigma}^2_{Cq_g} = \frac{s_I^2}{n_I} + \frac{s_J^2}{n_I n_J} + \frac{s_K^2}{n_I n_J n_K} + \frac{s_L^2}{n_I n_J n_K n_L}, \tag{2.4}$$

where s_I^2, s_J^2, s_K^2, s_L^2 are the variances of the subject, sample, RT, and qPCR levels, respectively, estimated from the pilot data. Additionally n_I is the number of subjects, n_J is the number of replicate samples from each subject, n_K is the number of replicate RTs from each sample, and n_L is the number of replicate qPCRs from each RT. By varying the n replicates at each level the $\hat{\sigma}^2_{Cq_g}$ can be changed. The optimal design is the one in which $\hat{\sigma}^2_{Cq_g}$ is minimised.

The inclusion of a financial cost into the calculation of the optimal design is trivial; the total cost of the experiment is

$$C_T = c_I n_I + c_J n_I n_J + c_K n_I n_J n_K + c_L n_I n_J n_K n_L, \tag{2.5}$$

where c_I, c_J, c_K, and c_L are the costs of producing a subject, sample, RT, and qPCR.

2.2.3 Statistical power

The power of a statistical test is the probability of rejecting the null hypothesis, given that the null is false and the alternative hypothesis is true. Power is simply a restatement of the Type II error rate, β, of falsely accepting a null hypothesis; power $= (1 - \beta)$.

The power depends on two factors; the significance criterion and the effect size. The significance criterion, α, is the Type I error rate of falsely rejecting a null hypothesis and must be specified before the power can be calculated. The α is often referred to as the rate of false-positives and the β as the rate of false-negatives. The α and β symbols used in terms of the significance criterion and power bear no relation to the α_{gi} and β_{gij} introduced in Eq.(2.1). For the purposes of this method only two classes of power calculation are considered; that used for the testing of the average expression of a single subpopulation in terms of a difference from a pre-specified value, and that used for the testing of the means of two subpopulations in terms of the difference from each other.

The effect size, d_1, in the case of a comparison of the mean expression of a single subpopulation from some pre-determined value, c, is simply

$$d_1 = (m_A - c)/\sigma_A \tag{2.6}$$

where m_A and σ_A are the mean and standard deviation of the subpopulation,

respectively, and correspond directly to μ_g in Eq.(2.1) and $\hat{\sigma}^2_{Cq_g}$ in Eq.(2.4) .

The effect size, d_2, in the case of a comparison of two subpopulations with unequal variances is defined as the difference between the means of the subpopulations divided by the precision of the measurement of each, thus

$$d_2 = \frac{|m_A - m_B|}{[(\sigma_A^2 + \sigma_B^2)/2]^{1/2}} \tag{2.7}$$

where m_A and σ_A^2 are the mean and variance of one subpopulation and m_B and σ_B^2 are the mean and variance of the second subpopulation.

Given the number of samples in the subpopulation(s) and the desired significance criterion the power can either be determined from a table, as found in [137] for example, or calculated from the cumulative distribution function of the t-distribution. N.B. a compensation is required when using a table to find the power of a test using a single subpopulation such that the effect size, d_1, should be multiplied by $\sqrt{2}$ to compensate for the fact that the c is a hypothetical population parameter without any associated sampling error.

2.2.4 Software implementation

Here we present a software implementation of this model. The software has been designed specifically for use with small pilot datasets where the variance structure of the experiment design is to be estimated. The software is written in Java and standard Windows / Mac OS X installers have been provided for native operating-system integration, while the '.jar' executable is available for Linux users. The majority of the software construction was performed using the Eclipse IDE, however most graphical user interface design and implementation was undertaken in the NetBeans IDE. At the user-level, *powerNest* has been designed to be lightweight and intuitive, requiring minimal computational or human resources for installation and operation, while at lower-levels the software employs efficient custom data structures that reflect the hierarchical nature of the experiment data. Initially based on user-modifiable default parameters, calculations of variances and power are performed in real-time while the user modifies the pilot-experiment design and updates displayed results as required. A detailed tutorial on the operation of the software, also discussed in the remainder of this subsection, is available at *www.powernest.net/PowerNest/Guide.html*.

Cq data can be entered into the software as MS Excel spreadsheets or plain text files. The Cq values must be allocated to the correct position in the experiment design hierarchy so that the software can determine to which subject, sample, and RT replicate each given data point belongs. This allocation can be performed manually in the software or be pre-specified in the input data file. In the case of the latter, a template file

is available; the use of which enables the software to automatically parse the experiment design. The user-interface for data input, design specification, and results analysis is shown in Figure 2.2.

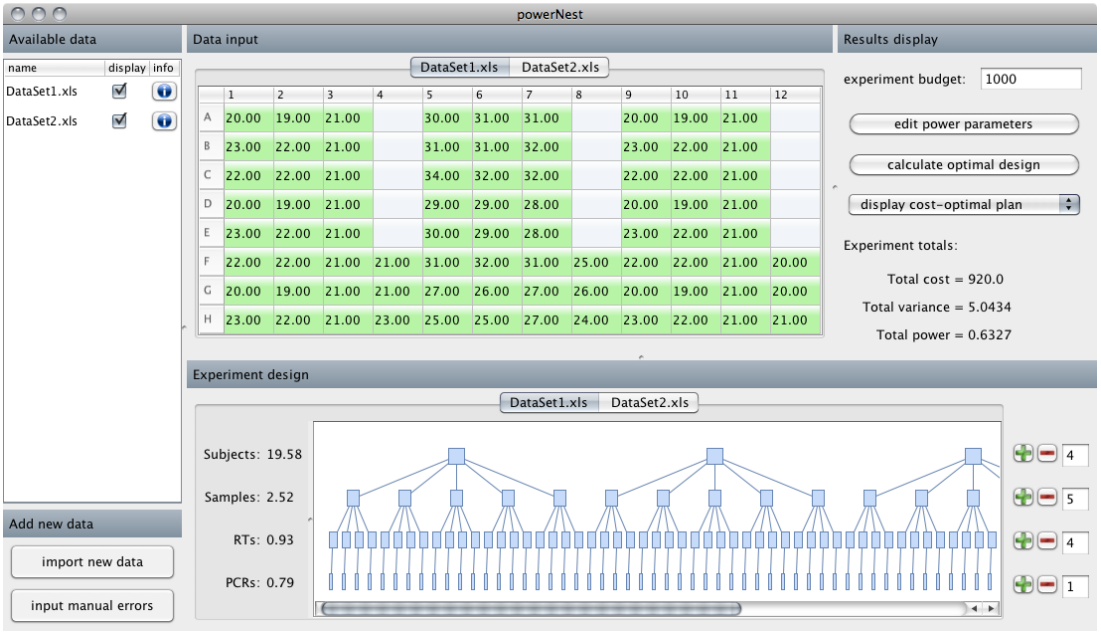


Figure 2.2: The main interface of the *powerNest* software. Cq data for a number of subject groups from a pilot study can be entered and grouped to provide estimates of the variance components of the experiment. Subjects within each group are assumed to be biological replicates; qPCRs from the same RT, RTs from the same sample, and samples from the same subject are assumed to be technical replicates.

Once the data have been allocated to their position in the experimental hierarchy, the nested-ANOVA automatically performs the analysis to determine estimates for the variance components of each of the four levels. These components are reported in terms of the relative, fractional contributions to the total variance in the data as detailed in Eq.(2.3). In the situation where experiment data are unavailable, a facility to manually input error estimates is also provided for each level. The results generated by using this facility must be interpreted with caution, depending on the researchers confidence about the quality of the input variance estimations.

Once the data are allocated to their correct position within the design of the pilot experiment the user is provided the opportunity to modify the design and sampling plan by adding and removing biological and technical replicates at each level of the design. Using the measured variance structure, error estimates and statistical power are automatically calculated and displayed for each of the modified designs. A facility

is also provided for inputting the approximate financial cost of performing a single replicate for each level. If this information is available, the software will display the total cost of each design as well as the expected total error.

Given the variance and costing information, the software is capable of determining an experiment design that minimises the total variance for a specified financial cost. This is achieved through an implementation of Eqs.(2.4) and (2.5). The user can choose from various designs such as those optimised for cost-performance, for the overall minimisation of biological and technical error, or for the maximisation of statistical power. Calculations involved in producing these suggested designs are deliberately limited, so as to remain computationally tractable, by a mandatory maximum experiment budget and can optionally be further constrained by limiting the maximum number of subjects, samples, RTs, or qPCRs to allow in the final design. These options are made available as simple user-interface elements and provide the user a rapid means of creating effective experiment designs.

2.2.5 Power calculation

For a single dataset from a single treatment group, the power of different experiment designs can be estimated in terms of the difference of the mean of the given data compared to a pre-specified value. The population variance is estimated either from the variance of the input data, or by manual estimate.

Data can be entered for multiple treatment groups such that the entire experiment design can be optimised based on the observed variances. Given this information, the software provides an automatic power calculation such that the statistical resolution of the assay for the desired contrast can be maximised before the experiment is performed. The automatic optimisation of the entire experiment design is capable of producing designs where the replicate structure of each treatment group is unique, enabling the overall error of the entire experiment to be minimised (i.e. different designs for each subject group depending on the result of the nested ANOVA).

The power is calculated based on the measured variance structure of the input data for the treatment group(s) using the effect size formulae defined in Eq.(2.6) or Eq.(2.7). For each design, the power is calculated using the number of biologically distinct observations (usually subjects, sometimes samples), the difference between the means of the treatment groups, and the precision of the measurements. The difference between the means can be either specified manually (preferred) or estimated from the data. The software can also plot a graph of the number of biologically distinct observations vs. estimated power, examples of which are illustrated in Figure 2.3.

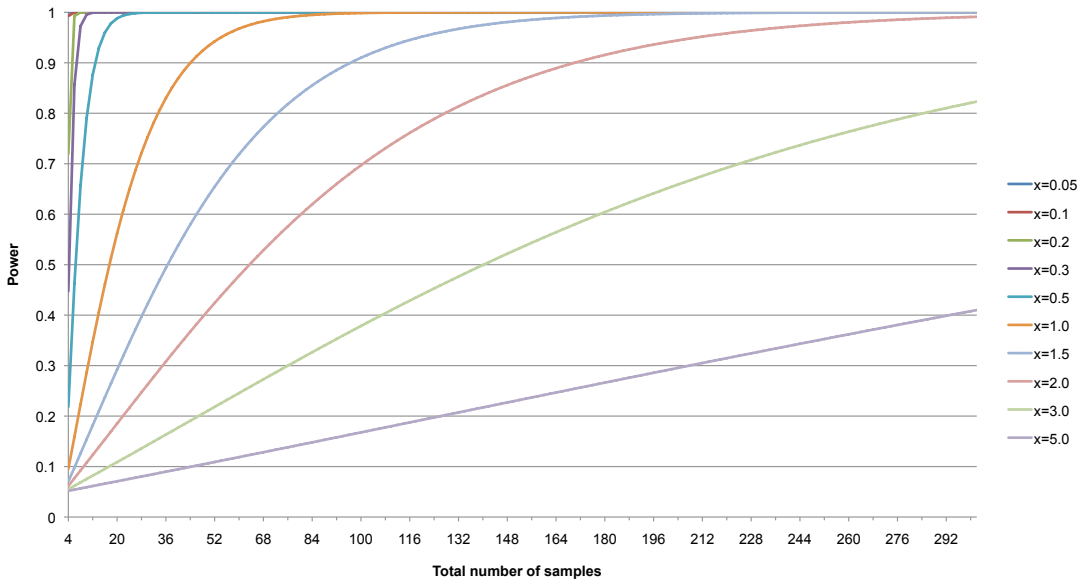


Figure 2.3: Example power-curves for a number of theoretical experiments at various measurement precisions at an alpha-level of 5%. The power is calculated for unpaired, two-tailed t-tests between two groups of samples of equal size and equal variance. For each curve, the variable x is defined as the fraction of the standard deviation of the two groups compared to the difference in the means of the two groups. For example x=0.5 corresponds to a standard deviation that is equal to half the difference in the means, x=1.0 corresponds to a standard deviation that is equal to the difference in the means, and x=3.0 corresponds to a standard deviation that is equal to three times the difference in the means.

2.2.6 Experimental application

Samples

This method and the software were first used in the analysis of several different types of biological material by Tichopad et. al. [144], in which the relative contributions of each of the processing levels to the total variance were estimated for bovine liver, blood, and culture samples for a number of different genes. The liver tissue was obtained from slaughtered heifers, as were the blood samples, in addition to cultures of adherent, growing IPI-2I cells from porcine ileum. For each sample we reverse-transcribed 500 ng of total RNA and amplified the cDNA by qPCR.

Results

The sampling plans for each of the sample types in this study were designed to include sufficient biological and technical replication to allow the estimation of the variance

components by the nested ANOVA, Figure 2.4(A). Here we extend this analysis to estimate variance components of the same data normalised to the reference gene, ActB, in each of the three tissues, Figure 2.4(B). We use these data to compare the measured variance structures before (Cq) and after (ΔCq) normalisation to the reference gene.

Prior to normalisation, the analysis of the liver tissue revealed substantial variation with an average total standard deviation that, in terms of the Cq, corresponds to a 2.6-fold variation between measurements. In blood, the noise arising from sampling and extraction was consistently small across all of the studied genes, both before and after normalisation to the reference gene, indicating that this step is very reproducible for such samples. The cell culture samples were found to exhibit the lowest overall confounding variation, attributable to the clonal nature of these cultures.

In all studied genes, with the exception of the low-expressed FGF7 in liver and IFNγ in blood, the magnitude of variance attributed to the RT step was reduced after normalisation. Excluding FGF7 and IFNγ, the estimated standard deviations at the RT step ranged between 0.18 - 0.46 cycles with a mean of 0.31 cycles in raw data, and were reduced to 0.03 - 0.25 cycles with a mean of 0.17 cycles following normalisation. The total standard deviations observed in blood and culture samples were only marginally affected by normalisation, while the total standard deviations of genes in liver (excluding FGF7) were dramatically reduced. The total standard deviation in FGF7 more than doubled following normalisation due to a large increase in the variance attributed to both the sampling and RT steps; the reason for this is unknown and with only a single observation we cannot speculate as to the significance of this result.

Many published reports have described the use of experimental protocols that perform only qPCR replicates. On the basis of the variance contributions we have estimated for the 3 studied sample types, we are able to evaluate the importance of qPCR replicates. Again excluding the low expressed genes, FGF7 and IFNγ, we found the standard deviations in raw data at the qPCR level to be 0.07 - 0.21 cycles, with a mean of 0.13 cycles; similar to previous findings [141]. We conclude that a qPCR standard deviation of 0.13 cycles is a good estimate for genes that are expressed at reasonable levels and assayed with a protocol that yields at least some 25 template copies per qPCR.

2.3 Concluding Remarks

The *powerNest* software application was specifically developed to implement the method presented in this article; it calculates the biological and technical variance components for a given dataset and can deliver cost-optimal, variance-minimising

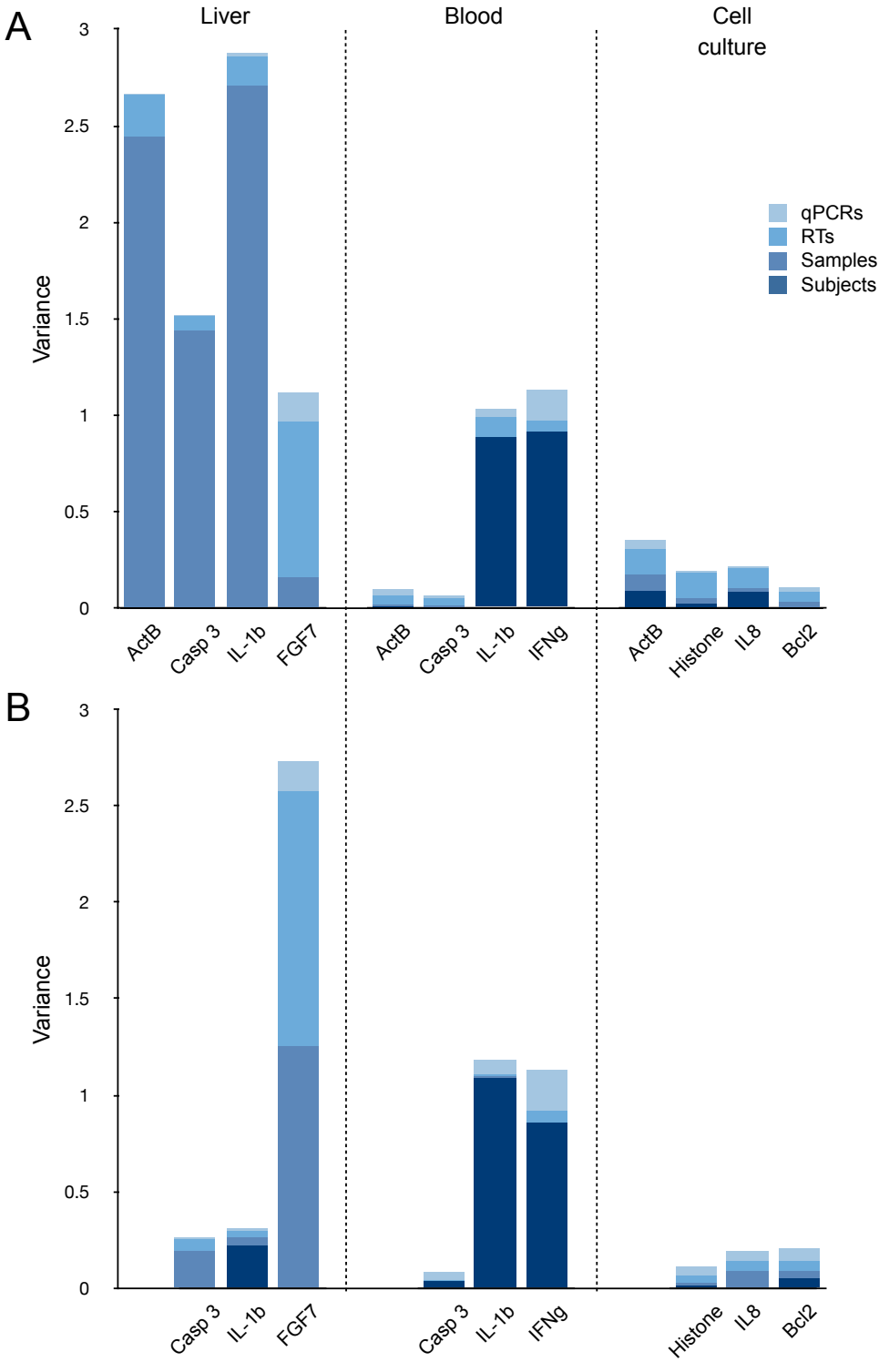


Figure 2.4: The estimated variance contribution of each of the four sampling levels to the overall variance in the measurements for several bovine tissues and genes. The top row of plots, **A**, illustrate the variances of raw Cq data while the second row, **B**, illustrate variances of ΔCq data after normalising to a reference gene- ActB.

experiment designs. Multiple datasets can be analysed simultaneously such that an estimate of statistical power can be calculated for a specified contrast between them. There are currently several published algorithms and software tools that address the analysis of gene expression with data generated by qPCR experiments; these include, among other things, different approaches to normalisation, the use of reference genes, and clustering of multiple targets and samples [146]. In addition, generalised software implementations of the nested-ANOVA and power calculations are also available [147, 148]. *powerNest*, however, represents the first dedicated tool to assist the researcher throughout the planning phase of an experiment and is available online at www.powernest.net.

General results for each of the preprocessing levels in the experiments described here highlight the importance of choosing the correct design for the specific environment of the experiment, such as the tissues and genes to be analysed. Across all of the tissues and genes analysed, the variance contribution from the qPCR step was only around 10% of the total and the contribution from the RT was found to exhibit about 2-times this variability, a result that is in agreement with earlier findings [149]. Along with the RT step, the variance of the qPCR replicates was found to be independent of the gene being assayed. We conclude that the use of technical replicates at the qPCR level have minimal impact on the precision of the estimated Cq value, in agreement with previous findings [144, 17]. In almost all observations, normalisation to the reference gene reduced the variance attributable to the RT step and the total variance was reduced in cases where the variance structure of the reference was similar to that of the gene of interest. The variability between sample replicates was found to be highly tissue-dependent and inconsistent estimates of the inter-subject variation in blood and culture tissues suggest that this variation may be gene-dependent.

It should be highlighted that in order for the technique described here to be valid it is essential that the subjects, samples, and pre-processing procedures used for the pilot are representative of those taken forward to the larger assay. It is obvious that the likelihood of the pilot being representative is increased through the use of larger numbers of biological and technical replicates; however, a sensible compromise must be made to limit the size and cost of the pilot study. We would generally recommend that, for the pilot to offer meaningful variance estimates, no fewer than three replicates are used at each level. In addition, although the use of technical replicates increases the statistical power of the assay by increasing the precision of the measurements, technical replicates are not independent and do not increase the number of biological observations of the given subpopulation.

When measurement is expensive and/or the individual measurements are very precise it is preferable to add biological replicates rather than technical replicates. In

conditions, exemplified by the bovine liver described above, where the dominant source of variability is between measurements rather than between the biological replicates, the use of technical replicates will be very effective in increasing precision. In general, however, the most effective means of increasing the power and validity of qPCR experiments is to increase the number of independent biological replicates randomly selected from within each subpopulation.

Chapter 3

Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles

Robert R. Kitchen, Vicky S. Sabine, Andrew H. Sims, E. Jane Macaskill, Lorna Renshaw, Jeremy S. Thomas, Jano I. van Hemert, J. Michael Dixon, John M. S. Bartlett.

BMC Genomics. 2010 Feb;11(1):134

Preface

The content of this chapter is exactly that presented in Kitchen *et al. BMC Genomics* 2010, except three heatmaps, which appeared in the supplementary material, that have been removed to improve readability. The presentation has been modified so as to conform to the formatting guidelines for a Ph.D. thesis chapter, but the content is unchanged from how it was presented in the original article. This original article, formatted as it appeared in *BMC Genomics*, can be downloaded free of charge from the publisher's website.

Both the original article and this chapter were written by myself and, bar a few recommendations by my fellow authors and reviewers, the structure and content is my own. I was not responsible for any of the biological processing of the samples, including patient selection, tumour biopsy, RNA preparation, and array hybridisation; these tasks were performed by Vicky Sabine, Jane Macaskill, Lorna Renshaw, Jeremy Thomas, Michael Dixon, and John Bartlett. I did, however, perform all of the data analyses reported in this chapter and in the original article- with the exception of Figure 3.7 which was produced by Andrew Sims.

Abstract

Background: Microarray technology is a popular means of producing whole genome transcriptional profiles, however high cost and scarcity of mRNA has led many studies to be conducted based on the analysis of single samples. We exploit the design of the Illumina platform, specifically multiple arrays on each chip, to evaluate intra-experiment technical variation using repeated hybridisations of universal human reference RNA (UHRR) and duplicate hybridisations of primary breast tumour samples from a clinical study.

Results: A clear batch-specific bias was detected in the measured expressions of both the UHRR and clinical samples. This bias was found to persist following standard microarray normalisation techniques. However, when mean-centering or empirical Bayes batch-correction methods (*ComBat*) were applied to the data, inter-batch variation in the UHRR and clinical samples were greatly reduced. Correlation between replicate UHRR samples improved by two orders of magnitude following batch-correction using *ComBat* (ranging from 0.9833–0.9991 to 0.9997–0.9999) and increased the consistency of the gene-lists from the duplicate clinical samples, from 11.6% in quantile normalised data to 66.4% in batch-corrected data. The use of UHRR as an inter-batch calibrator provided a small additional benefit when used in conjunction with *ComBat*, further increasing the agreement between the two gene-lists, up to 74.1%.

Conclusion: In the interests of practicalities and cost, these results suggest that single samples can generate reliable data, but only after careful compensation for technical bias in the experiment. We recommend that investigators appreciate the propensity for such variation in the design stages of a microarray experiment and that the use of suitable correction methods become routine during the statistical analysis of the data.

3.1 Introduction

DNA microarray technology has rapidly seduced scientists and clinicians with the ability to simultaneously measure the expression of tens of thousands of transcripts, enabling data-driven, holistic comparisons of groups or populations of cells, subtyping tissues, or predicting prognosis [150, 151]. However, as with any method, sound experimental design is essential to generate robust results from microarray experiments, particularly given the issues of high dimensionality [152]. Sufficient care must be taken to identify and correct for sources of experimental bias alongside a cautious interpretation of the importance of reported differentially expressed genes [74].

Efforts to promote the routine formalisation and control of all stages of the experimental workflow have seen success and are increasingly promoted by journals and microarray data repositories [74]. More recent work suggests the need for the inclusion of more detailed information concerning the statistical treatment of data in order for results to be independently validated post-publication [153, 154]. Such standardisation is essential to researchers wishing to re-analyse published data or combine multiple datasets in a meta-analysis. However the utility of these standards to the individual researcher gathering, analysing, and interpreting the data in the first instance is largely overlooked.

Despite all efforts towards standardisation, it is still not possible to account for all potential sources of variation in the experiment workflow; identical experiments performed at different sites have produced significantly different results [155, 118, 156]. Inconsistencies between results generated using different microarray platforms [118, 130, 157] or generations of array [126, 158] have been highlighted and multiplicative, systematic biases have been shown to be introduced at many stages of the experimental process, even when using a single array platform [126].

The common practice of hybridising samples with no technical replication (i.e. one replicate of each sample per experiment) is a result of the relatively high cost of arrays, the perceived improvement in array manufacturing quality, and the difficulties of obtaining sufficient amounts of high quality mRNA from some clinical samples. This practice is fundamentally reliant on the assumption that the intra-experiment variability is of a small enough magnitude not to undermine the power of the assay to resolve interesting biological differences that may exist between predefined groups of samples. There is, however, mounting evidence [118, 126, 159, 160, 161, 162, 163] to suggest that this assumption may be flawed and that the technical variation between replicate samples should not be ignored.

A large amount of effort has been expended in assessing the reliability, reproducibility, and compatibility of results generated by a number of array platforms

within and between laboratory sites. The microarray quality control (MAQC) project, a US Food and Drug Administration initiative [118], explored the intra- and inter-platform consistency of microarrays using two reference RNA samples (a universal human reference RNA (UHRR) from Stratagene comprised of high-quality RNA from a mixture of 10 different human cell-lines (including breast) and a human brain reference RNA from Ambion) and primary samples processed on six microarray platforms at three different sites. The results of the MAQC and other studies highlight the fact that, despite the generally good consensus between results, data generated from different platforms, in different laboratories, by different investigators can be negatively affected by dataset-wide batch variation in the reported expression levels [118, 130, 164]. Several methods that can remove these batch differences have been proposed, tested, and evaluated. Batch effects have been shown to be minimised with correction methods such as, singular value decomposition [129], distance weighted discrimination [130], mean-centring [126], and *ComBat* [127].

It is slowly becoming accepted that batch effects are to be expected when combining data generated across different labs, by different researchers, or using different platforms [118, 130, 157, 126, 158]. There is a strong motivation to integrate multiple studies for meta-analyses that have increased statistical power afforded by larger sample- sizes, which can help to overcome basic limitations such as the inherent heterogeneity between biological subjects. Combined datasets can swell to include thousands of tumours and have been shown to lead to improved results and consensus findings [126, 165, 166, 167, 168, 169].

Some researchers are now aware of bias arising due to analysis of samples at different sites or the use of different microarray platforms. The MAQC project [118], for example, was a multi-site and multi-platform comparison study, while others deal exclusively with the integration of data generated at geographically distributed locations. This study, to the best of our knowledge, is the first to assess the propensity for introduction of batch-processing effects at the same site and using the same protocol, making use of the multi-array Illumina BeadChip platform. We go further than the MAQC study by analysing both a commercial reference RNA and primary clinical material. This approach enabled us to demonstrate that it is possible to generate robust and reliable results, without the need for technical replication of starting RNA, but only when batch-processing effects are identified and suitably minimised. In this study we demonstrate compelling evidence for the existence of confounding batch-processing effects within a single experiment, using RNA prepared in the same laboratory, arrays hybridised and scanned at a single site, using a single protocol, and quantified on a single platform.

We investigated intra-experiment batch-processing variability on the Illumina BeadChip [34] platform, as multiple arrays on each chip allow an investigation of

intra- and inter-run variation. This was achieved through the hybridisation of a sample of UHRR to a single array on each chip along with duplicate preparations of cRNA from fresh frozen breast tumour samples that formed part of a recent clinical study [170]. Intra-experiment variation is common in other assays, such as quantitative RT-PCR (qPCR), where technical replicates and inter-plate calibrators are used to increase statistical resolution.

3.2 Results

3.2.1 Data quality

A qualitative measure of the performance of the BeadChips used in this study is provided by a measurement of the fraction of probes that are consistently called to be detected or undetected over all arrays. Analysis of detection consistency in the UHRR data in this study was comparable with the MAQC results [118] with 60–70% probes consistently called all-detected, and 80–90% genes consistently called as either all-detected or all- undetected, across all arrays in each run (data not shown). The coefficient of variation (CV) between and within the runs of the experiment was also consistent with the findings of the MAQC study, with a mean CV in quantile normalised data of around 7.5% (see Figure 3.1). The Illumina arrays used throughout the MAQC study were the Human-6 (48K v1.0) BeadChips, which differ from the Human-8 (24K v2.0) BeadChips used in this study in terms of the number of features represented. The Human-6 (v1.0) chips contain twice the number of probesets available on Human-8 (v2.0), however a large percentage of these additional probesets have been found to be unreliable [171] and are all contained within a completely separate strip on the chip leading to normalisation issues [160]. The high level of agreement in the observed CV and detection calls suggest any differences between the array versions at the probe-level are small.

3.2.2 Inter- and intra-run variation of the replicate UHRR samples

A clear batch-specific effect was observed in the raw data when the correlations of identical UHRR samples were assessed over all available pairs across the five runs processed on different days as illustrated in Figure 3.2. Generally, the level of correlation was high (>97%), however several clusters of samples were observed that corresponded to the batch in which the arrays were processed. In particular the samples in run 2 appeared to be very tightly correlated with each other but poorly correlated with samples in run 4 (Figure 3.3A). Quantile normalisation was found to have only a marginal improvement in the overall correlation of the samples and anomalies, such

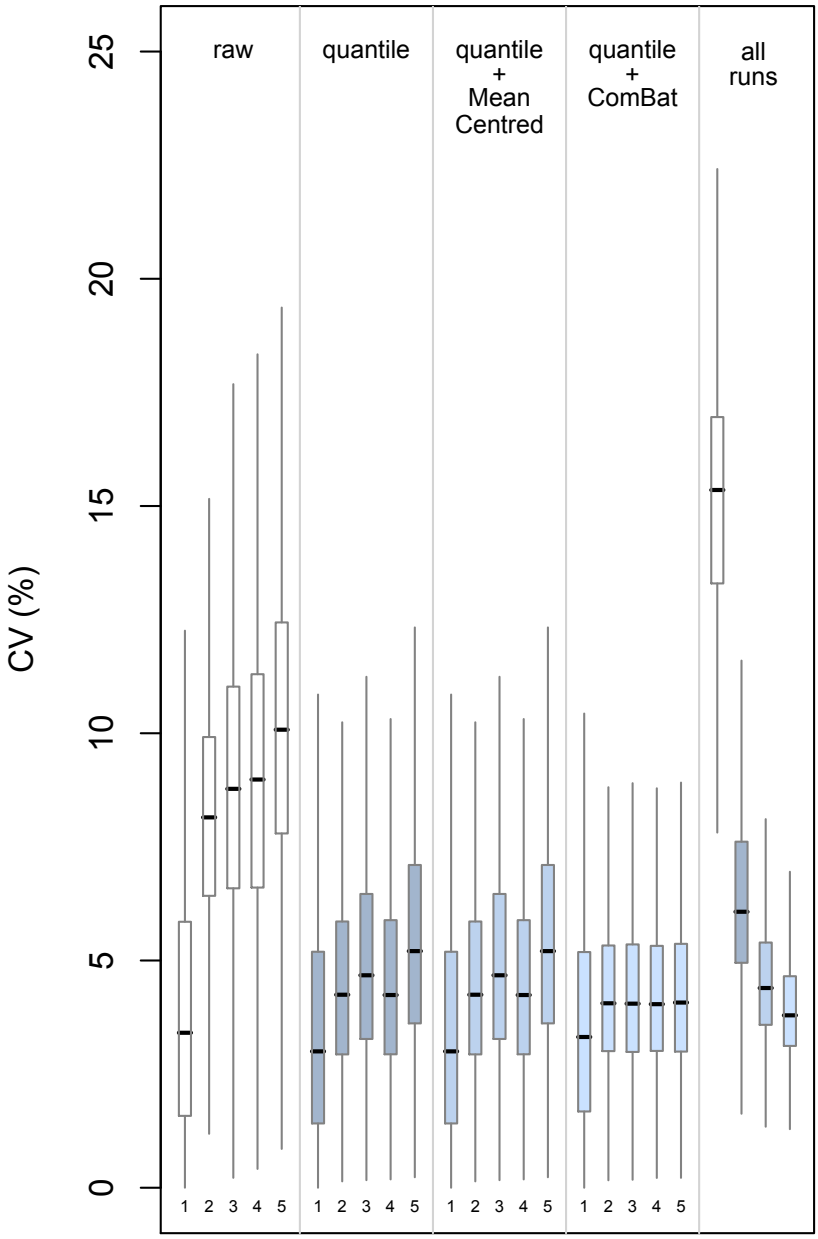


Figure 3.1: Coefficient of variation amongst replicate UHRR samples. From the left, the first four segments contain five box-plots illustrating the CV within each of the five runs; the four segments containing raw (white), quantile-normalised (dark-blue), mean-centred (lighter-blue), and *ComBat*-corrected (pale-blue) data respectively. All data were detection-filtered prior to analysis. The right-most segment shows the experiment-wide CV of the UHRR (coloured as the previous segments) calculated with no consideration of the individual runs.

as that between runs 2 and 4, were conserved (Figure 3.3B). Only on application of specialised batch-correction methods, such as mean-centring (Figure 3.3C) and *ComBat* [127] (Figure 3.3D), were these run-specific disparities shown to be substantially reduced. The correlations (calculated using Pearsons product-moment) for quantile normalised data ranged from 0.9833–0.9991, whereas following a *ComBat* correction this was increased to 0.9997–0.9999.

The probe-wise standard deviations of the raw expressions were found to be consistently small across the UHRR arrays (mean = 0.28). Using the nested analysis of variance described in methods, 60% (mean value) of the variability was due to that between runs and less than 40% to that within each run. The magnitude of the variation was marginally increased by detection-filtering (mean = 0.31), which would be expected due to the preferential filtering of probes with low signal. The application of quantile normalisation had a positive effect, decreasing the standard deviation to half that of the raw data. However both after detection filtering and quantile normalisation the relative contributions of the inter- and intra-run components to the total standard deviation remained approximately unchanged. Of a selection of other normalisation methods, loess, and cubic-spline performed similarly to quantile and all of these methods outperformed simple median normalisation (Figure 3.4). In all cases a further correction step is required after normalisation to correct for the batch effect.

Both mean-centring and *ComBat* reduced inter-run variation to such an extent that it could no longer be accurately detected by the nested-Anova method (Figure 3.5). The only observable difference between the two methods was that the *ComBat* corrected data also showed a slight reduction in the intra-run component of variation (Figure 3.5). The sequence in which the data were quantile-normalised and batch-corrected appeared to produce only marginal differences in the resulting variance components; as a result, all remaining corrections using mean-centring and *ComBat* were performed after quantile normalisation for consistency and to comply with the statistical assumptions of the latter [161].

The differences in measured expression between all combinations of pairs of UHRR samples that straddled the five runs (128 pairs) were calculated for raw, quantile-normalised, mean-centred, and *ComBat* corrected data (Figure 3.6A). The distribution of differences in the raw data did not resemble the expected form of a gaussian centred at the origin; instead it was skewed towards the positive (mean = 0.199). This was largely corrected after quantile normalisation and subsequent application of mean-centring and *ComBat* further narrowed the distribution reflecting the previously observed improvement in correlation. Similar improvements were observed in the differences between samples that were processed in the same run (25 pairs, Figure 3.6B). A full illustration of the intra-run pairwise differences can be found in Fig 3.7.

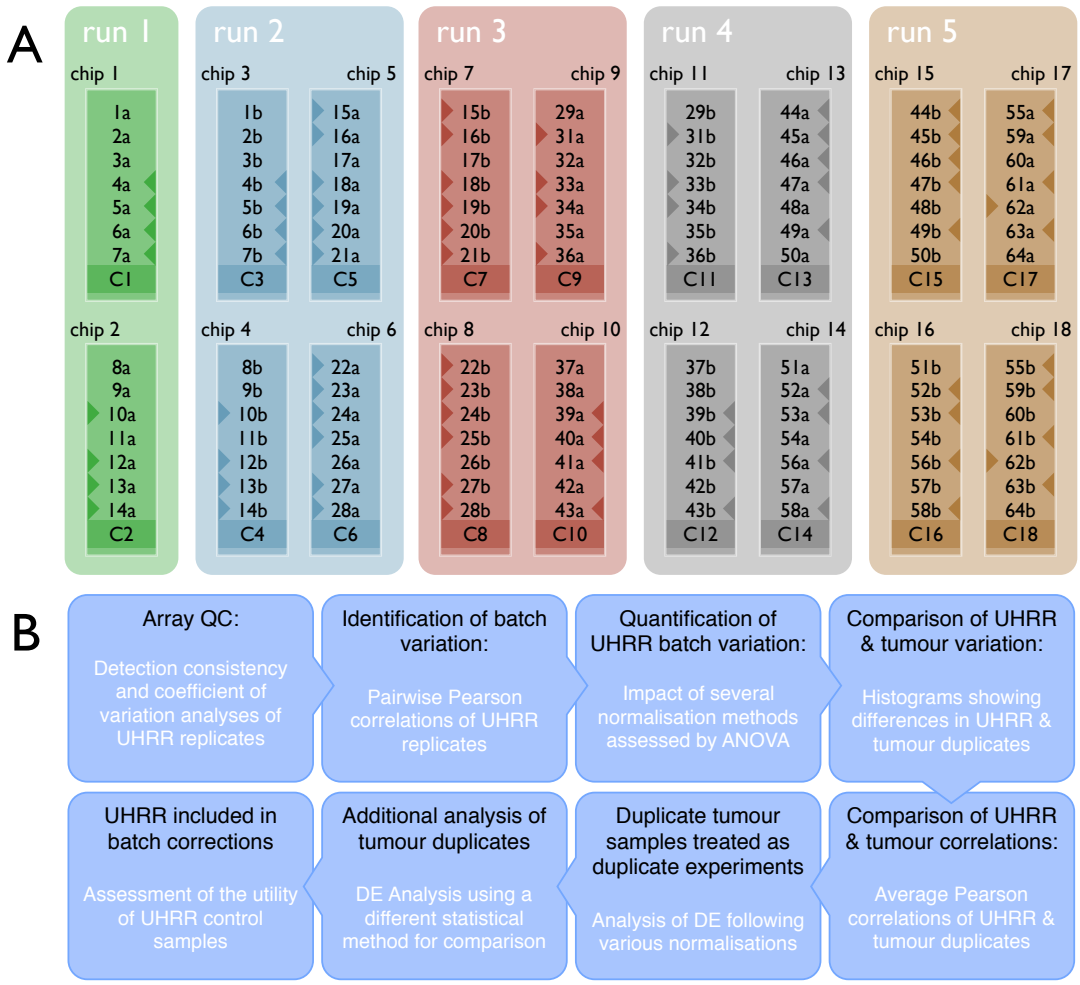


Figure 3.2: Layout of samples on the Illumina BeadChips and flowchart of the analysis approach. **A**: Illustration of the positions of samples on the 18 BeadChips, processed in five batches (also referred to as 'runs') corresponding to the five different days on which the samples were hybridised and scanned. UHRR samples are labelled as C1-18. Duplicate breast tumour clinical samples are labelled 'a' and 'b'. The pre- and post-treatment biopsy samples are identified by a triangle to the left and right of the sample IDs, respectively. **B**: Flowchart of analysis methods

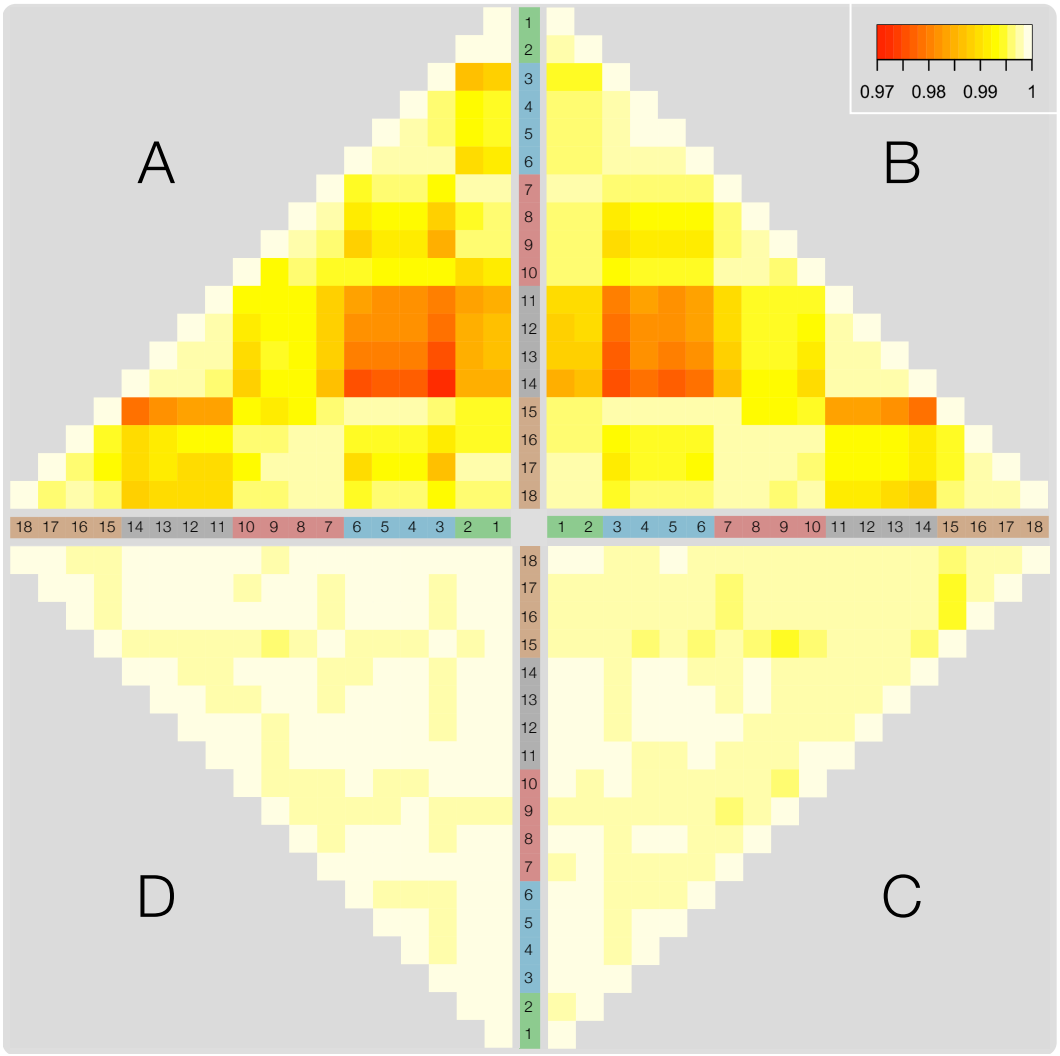


Figure 3.3: Intra and inter-run variation in UHRR samples: Pearson-correlations. Pairwise UHRR Pearson-correlation heatmaps highlight the batch differences, particularly between run 2 and run 4. Red cells correspond to ~97% correlation and white to 100% correlation. Batches and sample numbers are consistent with the colouring and labelling in Figure 1. All data were detection filtered, as described in methods. A = raw data; B = normalised; C = quantile normalised, plus mean-centring; D = quantile normalised, plus *ComBat*.

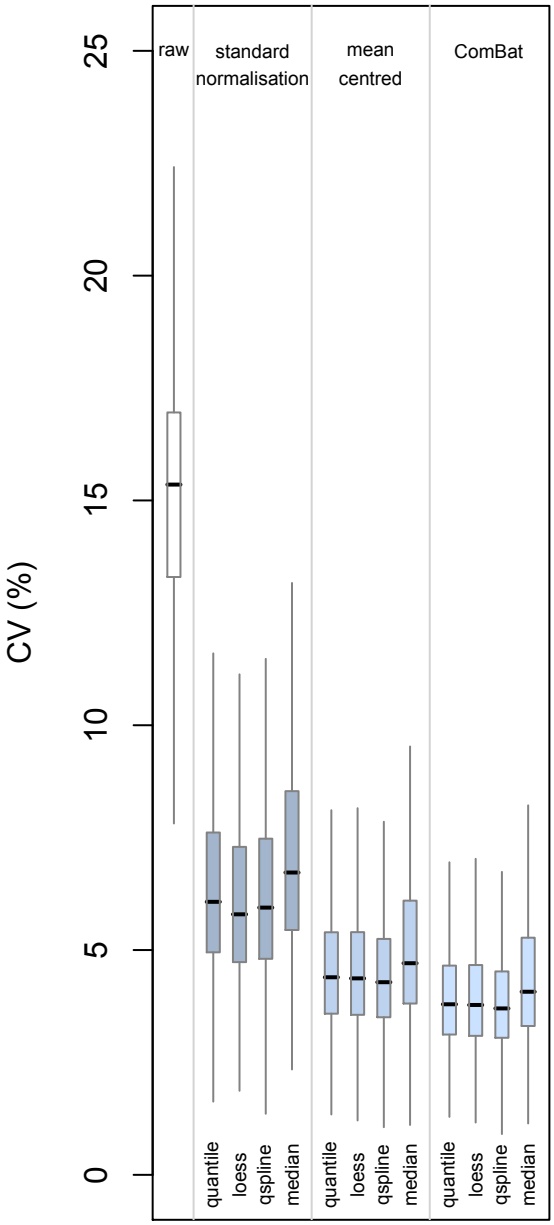


Figure 3.4: Coefficient of variation amongst replicate UHRR samples. This plot shows the experiment-wide CV of the UHRR samples. The left-most of the four main sections shows the CV of the raw (detection filtered) data, to the right of this is the CV after four popular normalisation algorithms; quantile, loess, cubic-spline (*qspline*), and median. The final two segments show the CV after batch-correcting each normalised dataset using either mean-centring or *ComBat*.

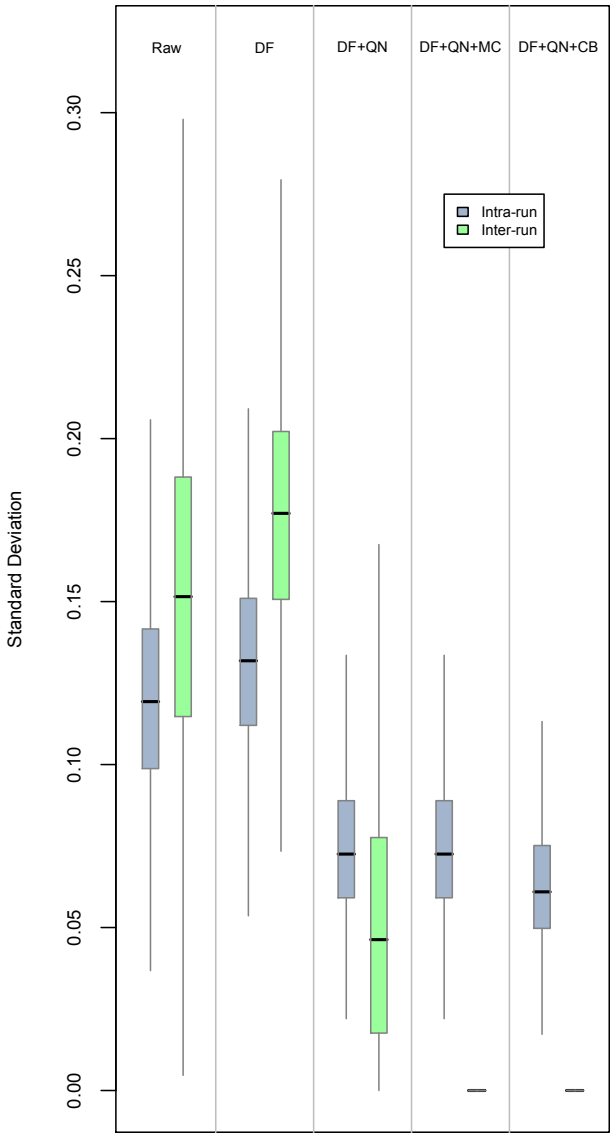


Figure 3.5: Intra and inter-run variation in UHRR samples: Nested-ANOVA. The results of a nested-ANOVA, quantifying the probe-wise components of variation corresponding to the within (blue) and between (green) batch variance. The model and calculation used are as described in methods. Effects on these standard deviations after detection-filtering (DF), quantile-normalisation (QN), mean-centring (MC), and *ComBat* (CB) are shown.

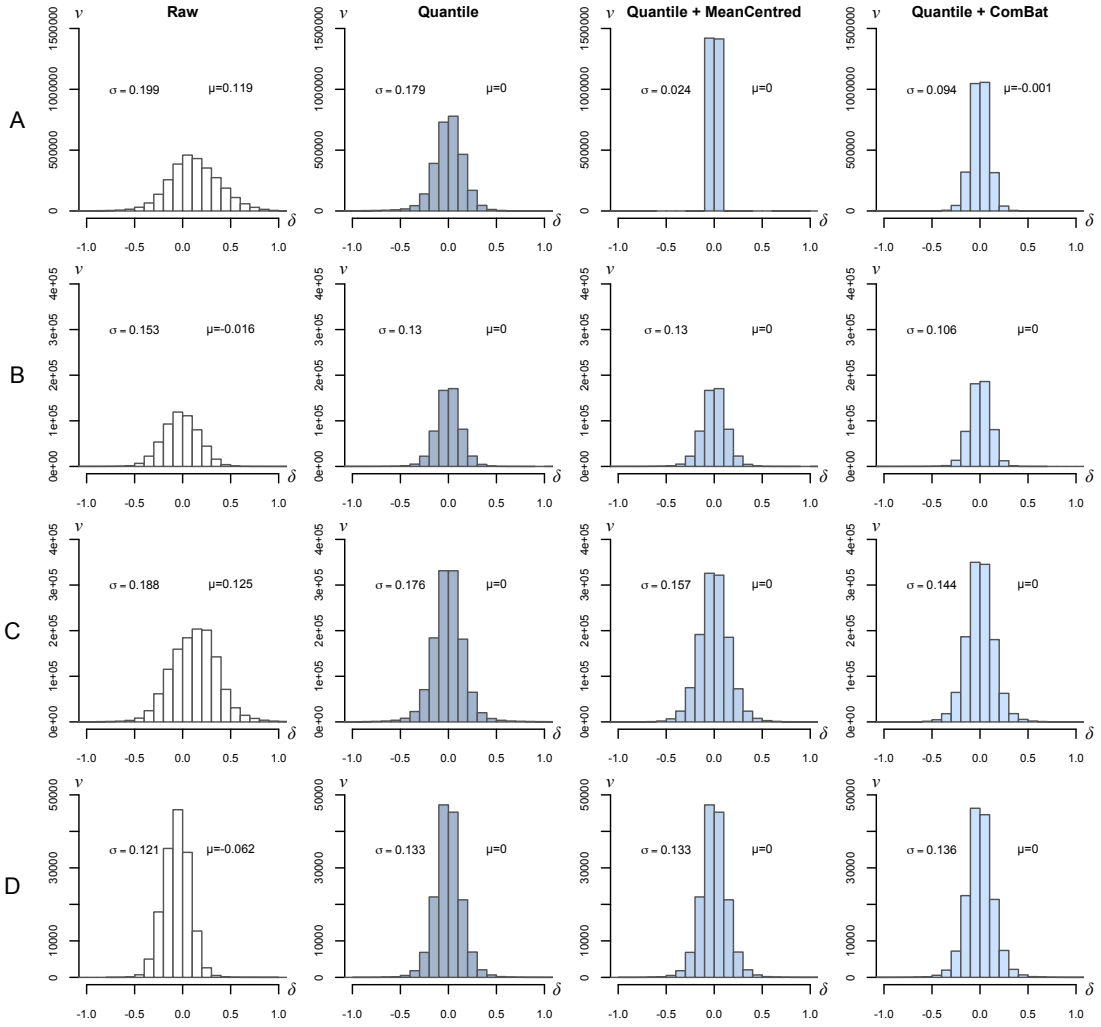


Figure 3.6: Distribution of the differences between replicate pairs of intra- and inter run intensity measurements. All possible combinations of differences between replicate pairs of UHRR controls and clinical samples were compared across the five runs. Axis labels represent the difference between duplicate samples (δ) on the x-axis, against frequency (ν) on the y-axis. Values on the left of each distribution represent the standard deviation and values on the right represent the mean of the measured differences. The four columns illustrate the effect of normalisation or batch correction on these differences. The four rows of plots illustrate both inter- and intra-run differences for both UHRR and tumour samples; row **A** contains inter-run differences calculated between the 128 pairs of UHRR samples; row **B** corresponds to intra-run differences between the 25 pairs of UHRR; row **C** is the inter-run differences in the 56 pairs of tumour samples; and row **D** contains data for the intra-run differences in 7 pairs of tumour samples in Run 5.

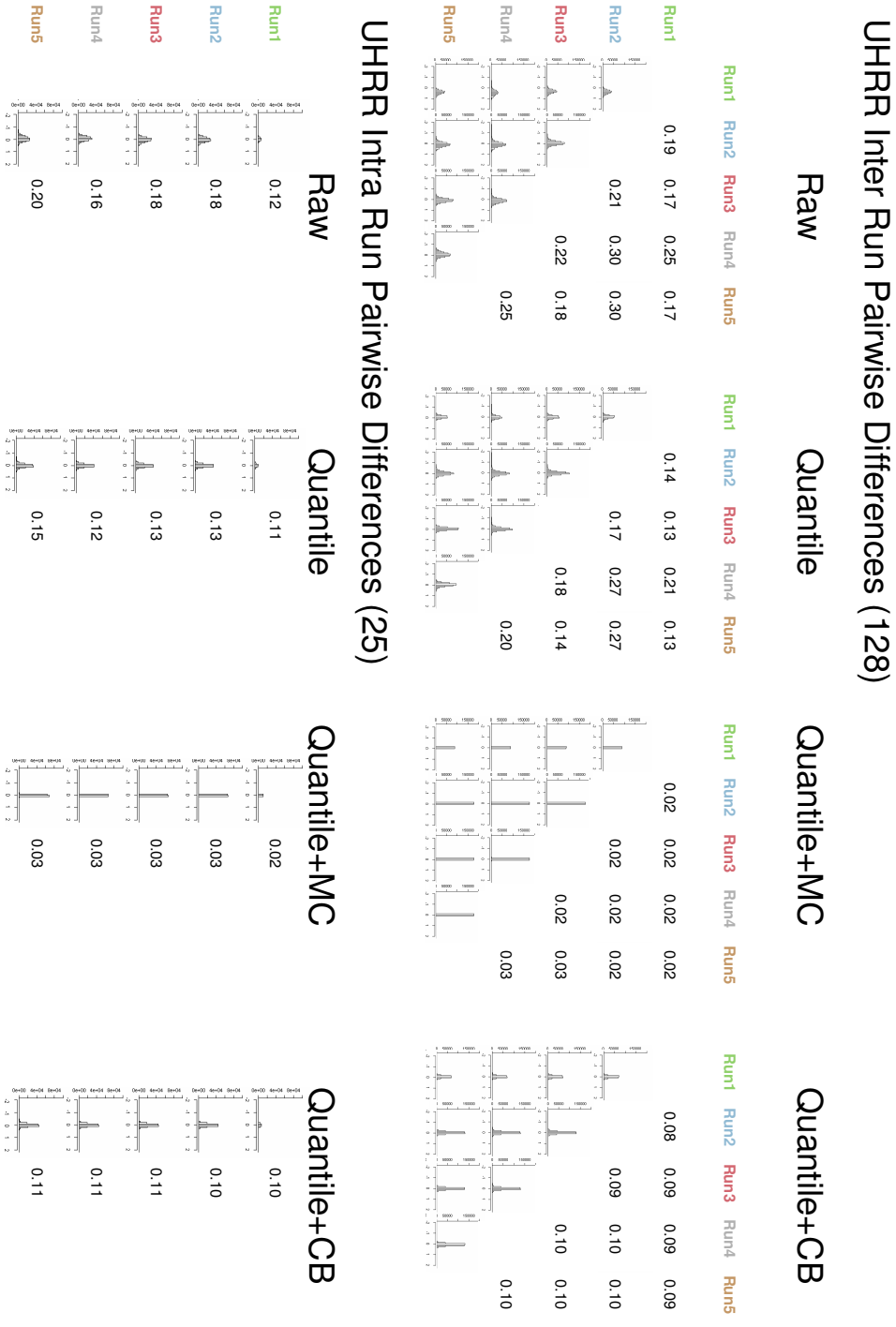


Figure 3.7: UHRR inter-run pairwise differences. Pairwise differences between each of the five runs calculated using UHRR samples for raw, quantile-normalised, mean-centred, and *ComBat*-corrected data.

3.2.3 Duplicate clinical breast-tumour samples

The sixty-three duplicate clinical samples provided a means to assess inter- and intra-run variation using samples more representative of those commonly analysed using microarray technology. The differences in the measured expressions between each of the duplicate pairs of the clinical samples that straddled the five runs (56 pairs) were calculated for raw, quantile-normalised, mean-centred, and *ComBat* corrected data (Figure 3.6C). As with the UHRR samples, moderate differences were observed between the raw expressions of duplicate hybridisations and quantile normalisation was found to reduce, but not eliminate, the differences between the duplicate samples. The distributions are similar to UHRR samples, although the raw data showed a slight negative skew that was again successively improved following quantile-normalisation, mean-centring, and *ComBat*, respectively. For completeness, intra-run distribution of differences between the duplicate samples was assessed for the seven pairs of samples in run five (Figure 3.6D).

Pearson's product-moments were calculated to assess the correlation between the duplicate samples. As with the UHRR the clinical samples were generally very highly correlated (>98%), although the samples on BeadChips 13/15 and 14/16 were found to be less similar than the others (Figure 3.8); this is consistent with the effect observed in run 4 using the UHRR. Batch correction by either mean-centring or *ComBat* increased the correlation for all samples except for two arrays on BeadChips 1/3 in the first run and all arrays on BeadChips 17/18 in the final run.

3.2.4 Comparing duplicate tumour samples as a repeated dataset to assess reproducibility of gene-lists

Of the 63 duplicate, paired clinical samples obtained from matched-biopsies before and after treatment with the mTOR inhibitor RAD001, 42 were of sufficient quality to be used in an analysis to reveal differentially expressed genes [170]. Using these samples we further assessed the impact of the intra-experiment variation in terms of the differences between lists of differentially expressed genes reported by each half of the duplicate samples. The hybridisation plan for the 21 pairs of pre- and post-treatment samples in each duplicate-group is illustrated in Figure 3.2; in the figure, triangles to the left of the sample represent pre-treatment samples and triangles to the right represent post-treatment samples. The first hybridisation of each duplicate sample is represented by a trailing 'a' and the second represented by a trailing 'b'.

The 'A' and 'B' duplicate sample groups, containing the 'a' and 'b' hybridisations of each sample, respectively, were considered as two completely independent datasets (as they were processed on completely separate BeadChips) in order to assess the extent

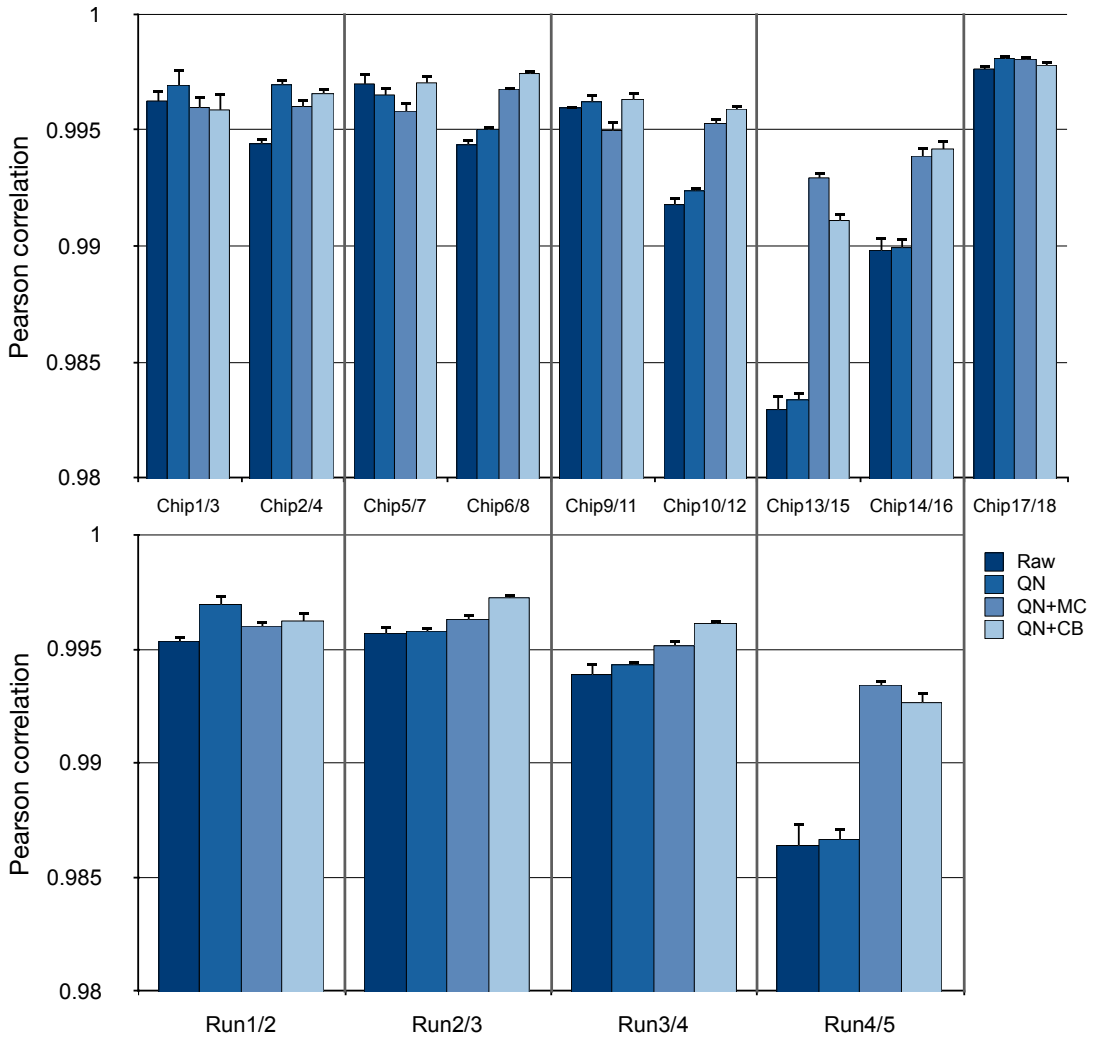


Figure 3.8: Intra and Inter-run comparisons of clinical duplicates. Mean Pearson-correlations between replicate pairs of tumour samples (A and B) on different chips and runs. Colours denote the four different data types; raw, quantile normalised (QN), quantile normalised then mean centred (QN+MC), and quantile normalised then *ComBat* corrected (QN+CB). Expressions were generally highly correlated except in the chips straddling runs 4 and 5. *ComBat* is able to correct for a significant amount of this difference. Error bars represent the standard error.

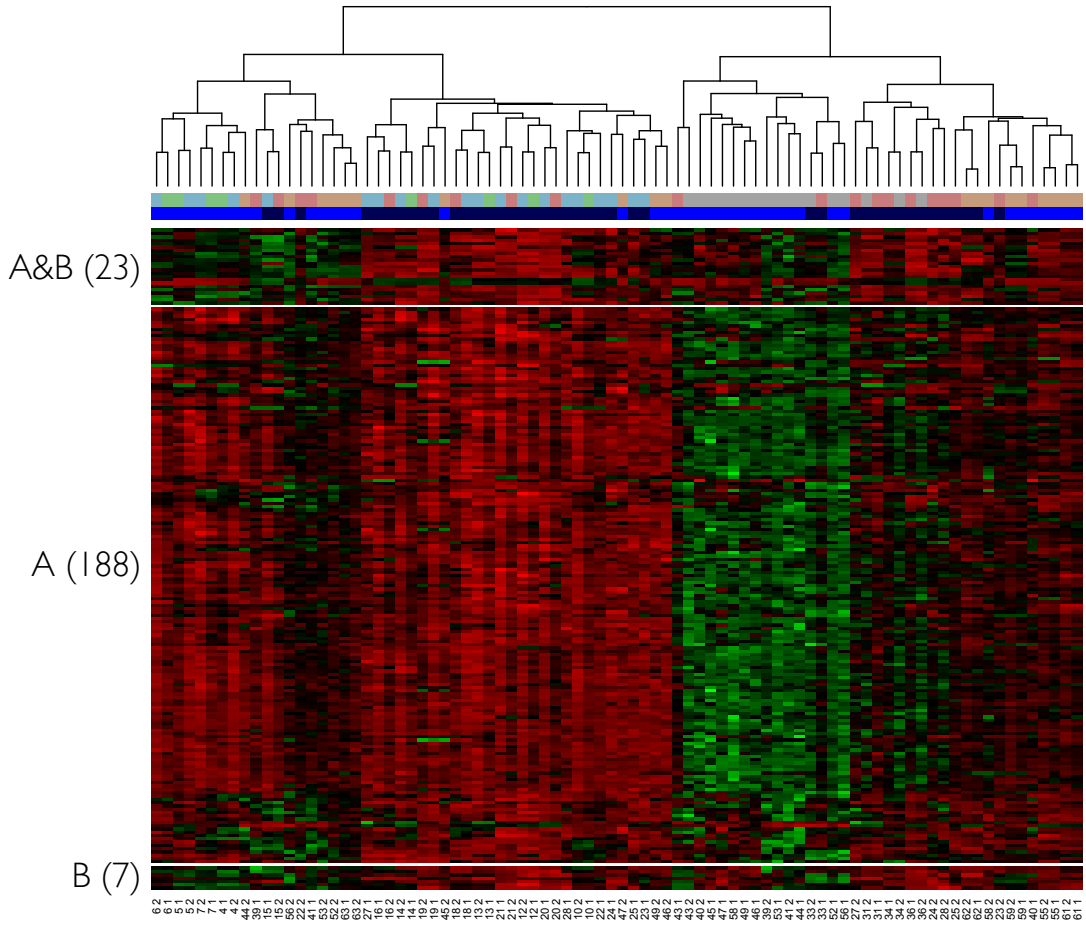
to which run-specific processing bias can influence the identification of differentially expressed genes. These datasets were independently filtered by detection calls, quantile-normalised, and, where stated, batch corrected by mean-centering or *ComBat* before generating lists of differentially expressed genes. Two BioConductor packages, *limma* and *siggenes*, were used to perform the statistical analyses (see methods).

Using the same stringency in the assessment of differential expression (fold-change ±1.5, adjusted p-value of 0.01) and using quantile normalised data, many more probes were found to be differentially expressed between pre- and post-treatment samples in sample group A (192) than in group B (30). Following batch correction with *ComBat* the number of differentially expressed genes identified in the two groups was more consistent (260 and 211) and the overlap, in terms of probes reported in both groups, increased from just 11.6% to 66.4%, however the use of mean-centred data only increased the overlap marginally to 15.2% (Table 3.1; Figures 3.9 and 3.10)

		A	B	A&B overlap	consensus(%)
limma	QN	192	30	23	11.6
	MC	225	222	59	15.2
	CB	260	211	188	66.4
SAM	QN	214	40	30	13.4
	MC	240	238	65	15.7
	CB	265	218	193	66.6
limma + UHRR	QN	205	31	24	11.3
	MC	8	92	7	7.5
	CB	144	119	112	74.2
SAM + UHRR	QN	224	42	32	13.7
	MC	17	100	12	11.4
	CB	149	125	117	74.5

Table 3.1: Summary of comparing the duplicate tumour samples as a repeated dataset (A and B) to assess the reproducibility of gene-lists. Differentially expressed genes were identified using *limma* and *SAM* as described in the text with quantile-normalisation (QN), mean-centring (MC), and *ComBat* (CB). The UHRR was used as an inter-batch calibrator.

Similar results were seen with less stringent criteria (fold-change ±1.2), which consequently led to larger numbers of probes, but similar proportions of overlapping probes were reported (data not shown). The analysis was repeated using significance analysis of microarrays (*SAM*) at a predicted false discovery rate of 5% and generated very similar results to those obtained using *limma*, increasing the overlap between groups of samples from just 13.4% in quantile-normalised data to 66.6% following *ComBat* batch-correction. See Table 3.1 for a full summary of these results. The


A&B (23)
A (188)
B (7)

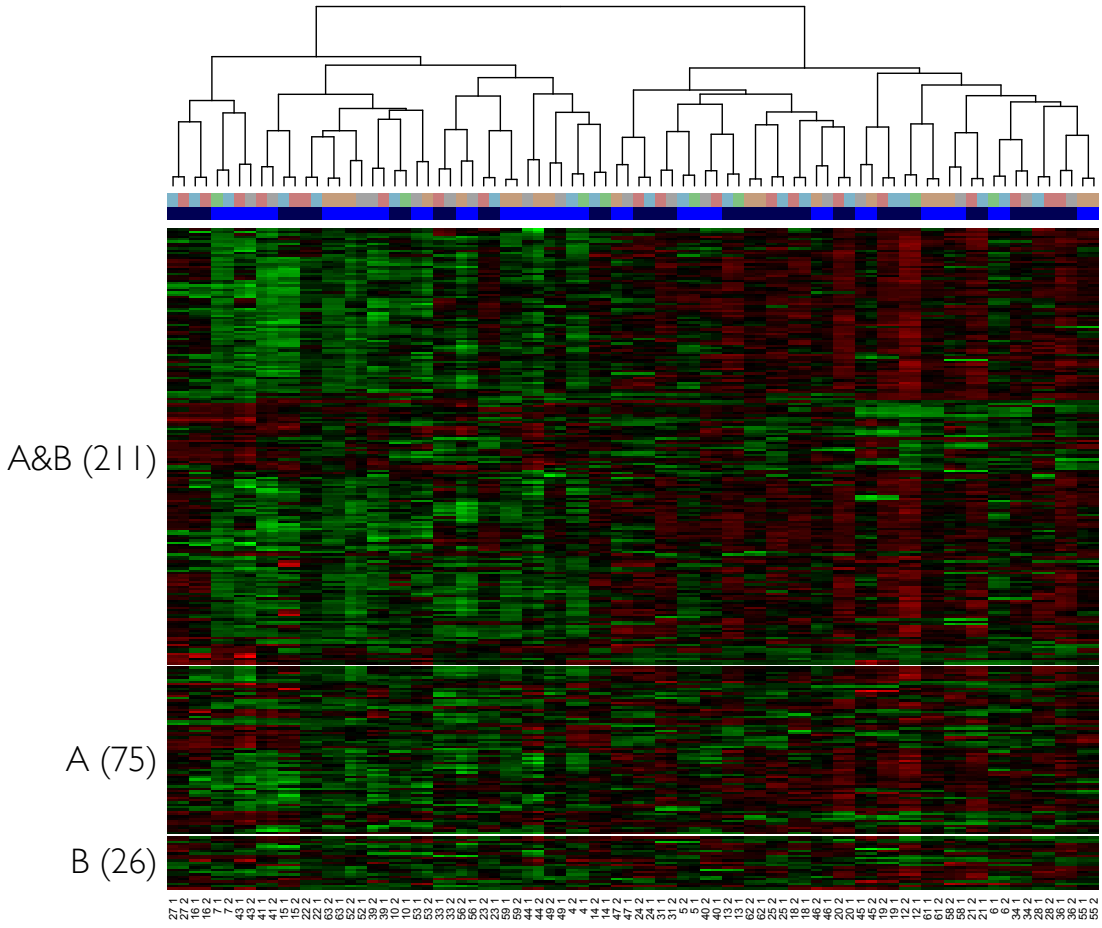


Figure 3.9: Differentially expressed genes with duplicates treated as separate datasets. Heatmaps of genes found to be differentially expressed in each of the A and B replicate datasets of samples and the overlap after quantile normalisation (top) and *ComBat* batch-correction (bottom). The batch in which each sample was present is denoted by bar beneath the dendrogram, in which the run-colours are consistent with those in Figure 3.2, and the sample-type is illustrated by the blue bar (light=post-treatment, dark=pre-treatment). The numbers of probes differentially expressed in both A and B ('A&B') or 'A' only and 'B' only are shown in brackets. Sample clustering (by complete linkage) in each heatmap was determined by only those probes in the 'A&B' group.

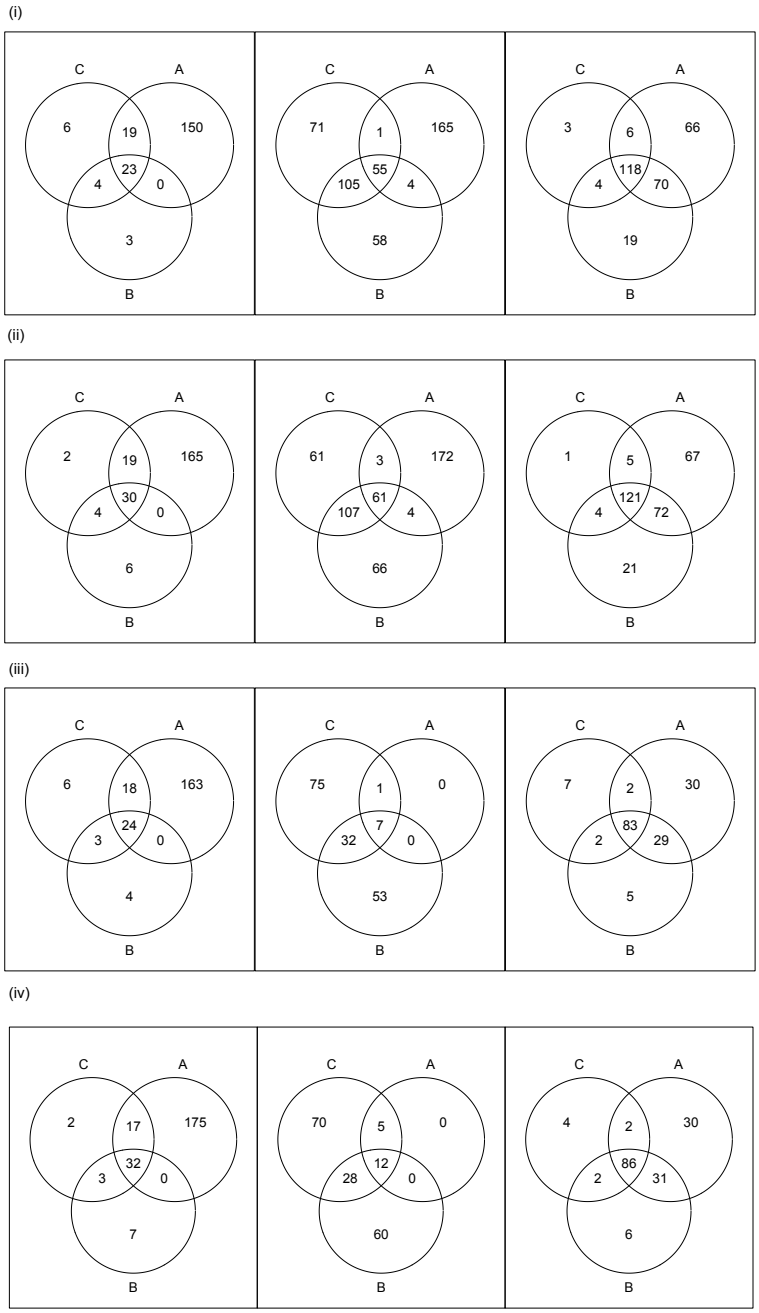


Figure 3.10: Numbers of genes reported to be differentially expressed after standard analysis (quantile normalisation) (left), after a standard analysis with mean-centring (middle), and after a standard analysis augmented with the *ComBat* batch correction (right). A and B refer to the results from independent analyses of the duplicate sample groups while C refers to the results from the pooled duplicate samples. The rows of Venn diagrams illustrate the results with (i) *limma*, (ii) *SAM*, (iii) *limma* using UHRR, and (iv) *SAM* using UHRR.

heatmaps in Figure 3.9 also highlight how all the pairs of duplicate samples cluster together following *ComBat* correction and the clustering is far less affected by processing runs. The dependency between the choice of *ComBat*, mean-centring, or quantile normalisation on the number of genes identified as differentially expressed in each replicate group was very strong in both the *limma* and *SAM* analyses ($\chi^2(2)$ p-value $<< 0.001$).

In addition to these independent analyses, the A and B groups were combined to create a third group of samples, 'C'. This group was analysed for differential expression in the same way and the results summarised in terms of the number of genes reported in any one, or any combination, of the three lists. The percentage of genes consistently reported by *limma* as differentially expressed in all three groups after *ComBat* correction was 41.3% compared to 11.2% after quantile normalisation alone and 12.0% after mean-centering (Figure 3.10). The percentage of genes identified in the pooled group C compared with those consistently reported in all three groups increased from 44.2% after quantile normalisation to 90.1% after *ComBat*. Again, very similar results were observed using *SAM* (Figure 3.10).

These analyses were repeated using the UHRR as inter-batch calibrator, designating it as a covariate in both the mean-centring and *ComBat* corrections. The inclusion of UHRR during quantile normalisation produced only a small difference in the number of differentially expressed genes identified in each of the three sample groups. However, the inclusion of the UHRR as a covariate in the mean-centring and *ComBat* corrections gave very different results. In both methods there was a large reduction in the total number of genes reported in each list, in terms of the consensus between the A and B groups, the agreement dropped to 7.5% following mean-centring, but increased to 74.1% after correction by *ComBat* (Table 3.1 and Figure 3.10). The dependency between choice of batch correction method and number of genes reported in either replicate group was stronger when UHRR was included in the correction in both the *limma* and *SAM* analyses ($\chi^2(2)$ p-value $<<< 0.001$).

3.3 Discussion

Batch-processing effects in microarray experiments are commonly encountered when combining datasets from different studies, different labs, or different technologies. In this study we have demonstrated that batch effects can arise within a single study, at a single lab, using a single technology and that these can have a significant impact on reported gene-lists.

The magnitude of the variation in the observed expression of replicate samples derived from the UHRR in this study is consistent with that reported in other studies

assessing the quality of microarray data, such as the MAQC [118]. We have also shown that the correlation of replicate UHRR samples is similar to that between duplicate pairs of samples derived from clinical breast-tissue biopsies and that this correlation is generally high. How- ever, when duplicate groups of clinical samples were independently analysed to identify differentially expressed genes the consistency of the resulting gene-lists was found to be very poor. The predicted false discovery rate of 5% using *SAM* was far lower than the observed proportion of genes that failed to be consistently reported over the replicate analyses (∼87% after quantile normalisation, ∼30% after *ComBat* correction). Whilst these two values are not directly equivalent, our results suggest that the predicted FDR may imply greater consistency than would be measured if duplicate samples are available. Specialised corrections for run-bias were more successful in reducing the magnitude of variability attributed to the inter-run batch effect in both UHRR and clinical samples. The reliability of results generated from the duplicate clinical samples was also greatly increased following batch-correction with a much greater proportion of genes consistently reported as differentially expressed in both sets of samples.

3.3.1 Use of single samples

There are many stages of sample-processing prior to conducting any gene-expression experiment and each is vulnerable to the introduction of systematic processing errors [126]. Opportunities to quantify this variation, prior to the microarray data analysis itself, are extremely limited and generally the only available option is an assessment of RNA quality. Other methods of quantifying gene expression, that are equally susceptible to the introduction of processing error, rely on the use of technical replicates to minimise confounding variation and maximise statistical resolution to the biological processes under investigation [144]. In this respect the routine practice of analysing each expression array sample as a singleton, regardless of the amount of RNA loaded, is an unusual scientific approach. Whilst BeadChip technology has a degree of built-in replication (approximately 30 randomly positioned beads, to which are attached ∼700,000 identical copies of a gene-specific probe [34]), this is no substitute for biological replicates, especially when a large degree of the observed error can be attributed to noise at the sample level, rather than at the probe level.

In the context of primary breast tumour samples, which have been repeatedly shown to have highly heterogeneous mRNA expression profiles, there is much greater variation between the RNA profiles from different individuals than within tumours [172]; either when comparing different tumour sections, biopsies and the tumour or FFPE and frozen [173], which effectively characterises the 'intrinsic profile' of subtype classification. On

this basis and the grounds of cost and scarcity of primary material it could be argued that replicates are unnecessary. However a lack of replicates limits the investigator in terms of their ability to assess whether the observed variation is of biological or technical origin and the extent to which it influences the resulting gene-lists. In this respect both biological and technical replicates are desirable to allow generated data to be screened for bias and batch-correction applied where appropriate. This is particularly important in the clinical setting if samples for large trials are processed in multiple labs.

Using the duplicate-experiment approach we were able to demonstrate that single samples can generate reliable data, particularly when batch correction is performed to minimise processing bias. However the genes reported to be differentially expressed in the pooled duplicate samples in group C were more robust in terms of their agreement with those identified in groups A and B, especially following batch-correction.

3.3.2 Use of UHRR controls

In addition to the technical replicates commonly used in other assays, in cases where the execution of the assay is split into several runs, it is very common for an inter-run calibrator to be used to quantify the variation introduced by the splitting of the experiment and to normalise for it. Despite the UHRR samples used in this study showing very similar variation and correlation to that previously reported, we found that the samples were of limited utility as predictors of the batch variation amongst the clinical samples. However the replicate UHRR samples were found to slightly improve the consensus between the results of the duplicate experiments when used in conjunction with the *ComBat* correction.

Although the UHRR has been reported to be useful as a standard for microarray experiments and suitable for monitoring the performance of genome-wide expression platforms [118, 156, 174], it has also been reported to not be a suitable representative as a normal sample for colon epithelial RNA [174]; similarly, the UHRR does not contain breast tumour RNA (only that from a breast cancer cell line among a pool). A more reliable control sample with which to improve the batch-correction might be provided by an mRNA sample more representative of that under investigation; in this case, a pool of tumour RNA rather than the UHRR. We found that the pre-treatment samples were good predictors of the batch variation amongst the post-treatment samples and so would likely make a better control (for normalisation) than the UHRR.

3.3.3 Experimental design

There is no reason to believe that the batch-processing effects observed here are limited to the Illumina BeadChip platform. Many previous investigations of other platforms

have postulated potential factors responsible for the introduction of processing errors in microarray experiments [126, 6, 120, 15]. Other experiments at our facility using the more recent Illumina Human HT-12 and Mouse Ref-8 BeadChips exhibit similar batch effects to those in this study; samples are observed to cluster preferentially with others processed in the same run, rather than by the biological differences between them, even after quantile normalisation (data not shown). Specialised batch-corrections appear to remove the bias, however without replicates such as those described in the current study, this cannot be fully evaluated.

Regardless of the platform chosen, it is clear that compensation for processing variation is beneficial and can only be achieved by incorporating the design of the experiment into the downstream data analyses. If all pre-treatment samples had been processed in one batch and all post-treatment samples in a second batch, it would not be possible to rule out confounding differences between treatment and batch processing. 'Real' differences due to the common variable of interest may have been partially or completely obscured by the batch effect. Design oversights of this type are beginning to be highlighted [153] and demonstrate the need to record the batches or processing runs in which data is generated. Some raw files contain metadata, such as the date in which they were generated, embedded within them. Acknowledgement and identification of the propensity for processing variation can be used to maximise the efficacy of batch-correction methods through a more informed design of the hybridisation-plan that includes, for example, randomisation and/or blocking of samples.

Our results support the notion that analysis of gene expression data should begin with an evaluation of batch effects. If the possibility of batch effects has been anticipated and confounding factors separated, then it should be possible to remove the bias to generate more robust results. As with other studies, ours is limited by the samples that were used in the evaluation of processing variation. We would have liked to test the applicability of our findings in other published datasets, however we were unable to find comparable datasets that include technical replicates and details of hybridisation 'batches' in the existing data repositories. In terms of cost and practicalities it is understandable why most researchers do not perform replicates in clinical studies, our results indeed suggest they may not be necessary; however providing a hybridisation plan along with the raw data, would make the processing of data more transparent.

3.4 Conclusions

In summary, intra-experiment bias can distort the findings of gene expression studies. Replicate samples were found to be beneficial in both the identification and reduction of processing bias and lead to increased consensus in reported gene-lists, especially

following specialised batch-corrections. We conclude that single samples can generate reliable data, although an appreciation for sources of intra-experiment variation during the design of the experiment is required to maximise the efficacy of specialised corrections in order to minimise susceptibility to potentially confounding intra-experiment batch-effects. Finally, based on the discrepancy between the lists of differentially expressed genes in each group of duplicate tumour samples, the observed rate of falsely-reported genes was consistently and significantly larger than that predicted by *SAM*. Therefore, based on the results of this study, a healthy degree of skepticism is advised when interpreting published results of microarray experiments that do not include validation by technical replication or, preferably, by another technique such as qPCR. In the absence of large numbers of biological replicates, it is our opinion that technical replication should be encouraged in order to provide robust, reliable, and credible expression-profiles.

3.5 Methods

3.5.1 Samples

In order to compare the consistency of gene expression profiles between and within processing runs a single sample of Universal Human Reference RNA (UHRR; Stratagene, Stockport, United Kingdom) was added to eighteen Illumina HumanRef-8 v2 Expression BeadChips. The remaining seven arrays on each chip were used to analyse the response to an mTOR inhibitor, Everolimus, in pre-operatively treated post-menopausal women with oestrogen receptor-positive breast cancer. From each extraction 100 ng RNA was amplified and biotinylated using an Illumina TotalPrep RNA Amplification Kit (Ambion) and quantified on a Bioanalyser 2100. 750 ng cRNA per sample was hybridised to Illumina HumanRef-8 v2 Expression BeadChips (Illumina, Cambridge, United Kingdom) using Whole-Genome Expression Direct Hybridisation kit (Illumina) and scanned with a BeadStation 500GX (Illumina). Full details of the sample biopsies taken at diagnosis and at surgery were as previously described [170]. The duplication of the clinical samples was performed after labelling and labelled samples were stored as per the manufacturers recommendations.

All raw gene expression files, clinical annotation and R scripts used to perform the analysis are publicly available from the caBIG supported Edinburgh Clinical Research Facility Data Repository *https://catissuesuite.ecmc.ed.ac.uk/caarray*.

3.5.2 Statistical methods

A summary work flow of the analysis approach is given in Figure 3.2. Gene expression changes were compared before and after RAD001 treatment and between responders and non-responders using Bioconductor [175] algorithms implemented in the statistical programming language, *R* [176]. Illumina probe profile expression data were normalised using quantile normalisation and corrected for batch processing effects using mean-centring [126] and *ComBat* [127]. Unless otherwise stated, the UHRR and breast tumour samples were normalised separately and the UHRR samples were not included as a covariate in the mean-centring or *ComBat* corrections. Genes differentially expressed between pre- and post-treatment samples were identified using *limma* [177] and *SAM* [67]. For the analysis using the *limma* package, genes were defined as being differentially expressed after satisfying a minimum fold-change of ± 1.5 and a maximum, Benjamini-Hochberg adjusted, p-value of 0.01. For the *SAM* analysis (using the *siggenes* package), the differentially expressed genes were selected at a maximum predicted false discovery rate of 5% and the same minimum fold-change of ± 1.5. Paired statistical tests were performed in both the *limma* and the *SAM* analyses. Hierarchical clustering of samples and probes for the creation of all heatmaps was performed using complete linkage and similarities calculated according to the method described in [178].

Data were filtered, where specified, using the detection confidence reported by Illumina's *BeadStudio* software- determined for each bead based on the expressions of internal control probes, local background intensity, and the uniformity of the reported intensity of the bead. The filtering was performed prior to normalisation such that probes with a detection confidence less than or equal to 80% in more than 25% of the samples were removed from further analysis.

We applied a linear additive model to UHRR expression data on the log-scale to estimate the inter- and intra-batch variance contributions. These contributions are assumed to be independent and randomly drawn from log-normal distributions. As all factors meet in unique combinations a nested, or hierarchical, variance model is individually applied for each gene such that the model of the measured expression, X_{ij}, of each probe is defined as

$$X_{ij} = \mu + A_i + \epsilon_{ij} \qquad (i = 1, ..., b; j = 1, ..., n) \tag{3.1}$$

where μ is the geometric-mean expression of the gene from the UHRR population, A_i is the random effect attributed to the ith batch, and ϵ_{ij} is the random measurement error attributed to the jth array in the ith batch. Finally, b is the total number of batches and n the number of replicate samples in the corresponding batch. The variance of any given observation, X_{ij}, is $\sigma_A^2 + \sigma^2$; these components represent the inter-batch and intra-

batch variance respectively. The estimation of σ_A^2 and σ^2 is performed independently for each gene as stated in [145].

Batch correction

The main difference between standard array normalisation and batch-correction is that the latter does not make the assumption that all probes are affected equally by batch-effects and, as such, performs individual adjustments on each probe across all samples.

Mean-centring, as used in the context of this chapter, is effectively N separate operations, where N is the number of probes/genes, in which each operation involves finding the mean expressions of samples within each of the batches and adjusting these sample's expressions such that these means share a common value. In the analyses described here, a bespoke method was created to perform the adjustments making use of the 'rowMeans' function in R and any batch in which there were observations of only one of the tumour sample-types (e.g. the batch containing only pre-treatment samples) were excluded from the adjustment. This procedure is not dissimilar to a single iteration of the 'median polish' algorithm in which a model of the row and column median-averages is found by iteratively finding and subtracting the median of the rows/columns from the data; these median values are recorded as row/column effects and, along with the sum of these effects, are then considered an estimate of large-scale variation in the data.

The ComBat approach, as discussed in detail in [127], is a little more involved than the mean-centring method but can be summarised in three stages:

1. **Data standardisation and global parameter estimation**

   First, using a method similar to autoscaling, the data over all genes are standardised so that they all have similar mean and variance. This step compensates for the different expressions and variations of the various probes/genes that would otherwise bias the batch effect estimates. Next the mean and variances of all samples in each batch, over all probes/genes, are estimated using a linear model and these parameters constitute the first prior-distribution.

2. **Local parameter estimation**

   Independently for each gene, the sample mean and variance is estimated for each batch and are used to estimate the parameters of the additive and multiplicative noise distributions using the method of moments; these constitute the second prior-distribution.

3. **Data adjustment**

Using the parameter estimates for these two prior distributions along with Bayes theorem, posterior distributions for each of the additive and multiplicative noise distributions are calculated; final values of these batch effect parameters are estimated as the expected values of the posterior distributions. This empirical Bayes procedure allows information from all genes to be used to estimate batch effects for each gene, providing more stable estimates than the standard sample mean and sample variance.

Chapter 4

Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments

Robert R. Kitchen, Vicky S. Sabine, Arthur A. Simen, J. Michael Dixon, John M. S. Bartlett, and Andrew H. Sims.
BMC Genomics. (submitted)

Preface

This chapter follows directly from the array experiment performed in the last chapter. Here is presented an analysis of a bespoke array dataset, using a more modern iteration of the Illumina BeadChip technology, in addition to two previously published sources of data; all of which complement and extend the data and the analyses presented in the previous chapter, as will be discussed.

The content of this chapter is also presented, in a condensed form, in Kitchen *et al. BMC Genomics* 2011 (submitted).

Both the submitted article and this chapter were written by myself and, bar a few recommendations by my fellow authors and reviewers, the structure and content is my own. As was also the case in the previous chapters, I was not responsible for any of the biological processing of the samples, including patient selection, tumour biopsy, RNA preparation, sample-pooling, and array hybridisation; these tasks were performed by Vicky Sabine, Michael Dixon, and John Bartlett. I did, however, perform all of the data analyses reported in this chapter and in the submitted article- with direction by Andrew Sims and Arthur Simen.

Abstract

Background: Systematic processing errors are extremely common in microarray experiments. When samples are analysed by the same technician, at the same time, and using the same technology, compensating for most obvious sources of technical variation is relatively straightforward. However in experiments such as those used to classify samples for clinical diagnostics, or any experiment in which the main effect under investigation is confounded with sources of systematic error, reliably resolving the effects is a non-trivial task. To better understand the importance of various factors in experimental-design, we assessed Illumina and Affymetrix datasets for technical variation between replicate sample-hybridisations.

Results: Variances estimated at the inter-experiment level between our HT-12 and Ref-8 BeadChips were more than twice as large (on average) in our study than the inter-lab variation in the microarray quality control (MAQC) dataset. Batch correction using established methods was very effective in removing systematic error attributed to the inter-experiment, -run, and -chip levels. Controls of both universal human reference RNA (UHRR) and pools of RNA derived from 'analysis' samples modelled technical variation well, with high correlations to duplicate pairs of tumour samples although the pools were significantly (on average approximately 4%) more highly correlated than the UHRRs over all probes. A lack of differential expression in the new tumour samples due to biological variation and relatively small numbers restricted more comprehensive analyses of the UHRR and pool as calibrators during batch correction. On correlating probe-wise standard deviation (SD) estimates with probe GC content, we found low-GC probes were significantly more variable to batch effects than probes with higher-GC content. This low-GC, high-SD correlation was significant in both our data and the MAQC Illumina dataset, but was not resolved on analysing a large set of biological replicates.

Conclusion: The primary source of systematic error in any given microarray experiment is unpredictable, however pools of sample RNA and commercial controls such as UHRR are effective in modelling the variation. We found that the pooled samples outperformed UHRR, better emulating the effects of systematic error, and would almost

certainly act as a more effective calibration sample during batch-correction. Probes with low GC-content are more vulnerable to systematic error but, although highly statistically significant amongst technical replicate samples, the magnitude of this variation is relatively small and is masked by biological variation. Blocking of samples from each sample-group both within experiment-runs and within each BeadChip are important to protect against technical noise in Illumina data and detailed meta-data should therefore be preserved for each array that includes the date and time of each hybridisation/scan. Diagnostic procedures such as PCA or SVA should be routinely performed prior to downstream array analyses for the detection of unexpected systematic error.

4.1 Introduction

Increased adoption of high-throughput, whole-genome gene expression analysis technologies has led to an increased focus on the reliability of the experimental measurements they produce. Several recent articles have provided substantial evidence of systematic effects influencing data from many technologies including microarray, second generation RNA-sequencing, and mass-spectrometric methods [119, 179]. Of these very different methods, the majority of evidence for such systematic effects is derived from microarray data due to their popularity, low cost, and relative ease by which large datasets can be generated [6, 150, 180].

Complete confounding of batches of array scans (even in the same laboratory, but at different times) with studied populations can be disastrous to the reputation of reported results. For example, a study by Spielman *et al.* [122] reported 26% of all genes differentially expressed between samples from European and Asian human populations. Re-analysis of these data however revealed that the arrays for each population were processed separately, between 1-3 years apart, and that after application of a standard batch-correction method [181] no genes remained significantly differentially expressed. This is a rather extreme example of a very common issue of vulnerability to confounding experiment noise in microarray study design. Several studies have attempted to assess reliability and consistency of array measurements and estimate potential sources of confounding experimental noise in an attempt to reduce these vulnerabilities [153, 159].

The microarray quality control (MAQC) project was set up to explore inter-platform and inter-laboratory consistency of microarray-derived gene expression datasets using two reference RNA samples [118], as well as consistency of differential expression estimates [182]. Both of these studies reported generally good consensus between replicate samples across technologies and laboratories. However the latter study found that filters used to determine collections of 'significant' probes/genes were less effective when based on the reported significance of the individual statistical tests compared to a filter based on the magnitude of the differential expression [183].

Analyses involving multiple laboratories, technologies, and staff are of interest when dealing with large collaborative investigations and in large meta-analyses, where multiple sources of existing data are collated with the intention of increasing the statistical power to detect subtle differences in expression. Direct comparison and integration of gene expression data through meta-analyses is a highly attractive option and resources such as NCBI GEO [184] and EBI Array Express [185] make the process of discovering potentially complementary datasets easier than ever. Unfortunately appraising the quality of publicly available data remains a non-trivial task and, even when the submission guidelines are followed correctly, it can be difficult to identify

technical effects that may lead to systematic bias in the data that can potentially compromise downstream statistical analyses.

Another, perhaps more common, situation requiring awareness of systematic batch variation is the use of arrays for clinical diagnostics, in which genome-wide expression patterns are mined to classify patient samples to one or more predefined disease phenotypes. It is generally the case that such applications have two distinct stages in which a training sample-set is used to build and tune a classification algorithm, before this system is used to classify or diagnose samples of unknown phenotype. These two stages are, in the clinical context, rarely performed at the same time or even using the same technology. It is therefore imperative that the technical variation between technologies, experiments, and runs is sufficiently small, or able to be reduced, such that unknown samples can be successfully and reliably classified.

We previously reported compelling evidence for the existence of confounding batch-processing effects within a single experiment, using RNA prepared in the same laboratory, arrays hybridised and scanned at a single site, using a single protocol, and quantified on a single platform [186]. In this moderate-scale experiment run on 18 Illumina Ref-8 BeadChips over the course of five days, replicate hybridisations of a Universal Human Reference RNA (UHRR) and duplicate hybridisations from fresh frozen pre- and post-treatment breast tumour samples [170]. The replicate UHRR samples (one per chip) detected a batch effect that was reflected in the duplicate pairs of tumour RNA and, following variance analysis, it was revealed that the processing batch (corresponding to the day on which the chips were processed at the core facility) was the main source of technical variation in the measured expressions. Importantly, although this variation was small, it had a profound effect on the internal consistency of duplicate analyses of differential expression in that only $\sim$10% of genes were consistently reported as significantly differentially expressed. This was only remedied following specific batch-correction using mean-centring, and further improved using *ComBat* [127] method, increasing the consistency to $\sim$70%.

4.1.1 Motivation and analysis plan

Given previously reported variation in various array technologies [70, 126, 186, 187, 188, 189], we wished to quantify the relative error introduced at various stages in the experimental process prior to array scanning; including the choice of array-version and study design, relevant in clinical diagnostics applications or replication studies. In addition, we wanted to assess different types of control sample in terms of their ability to model some of these systematic effects and we also considered specific properties of array probes in terms of correlation with such effects.

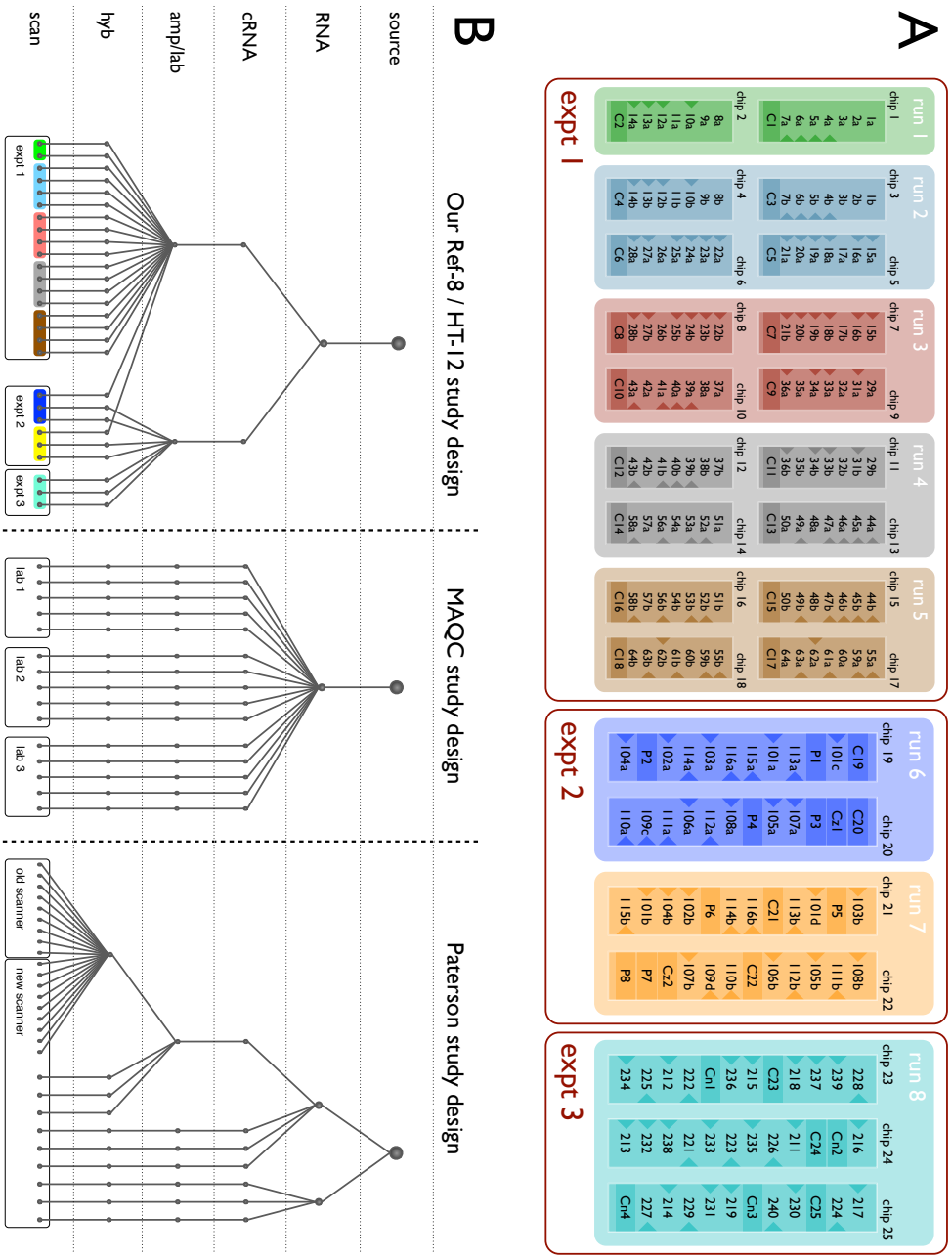


Figure 4.1: **A**: Illustration of our Illumina Ref-8 (experiment 1) and HT-12 (experiments 2 & 3) BeadChips, processed in eight batches (also referred to as 'runs') corresponding to the different days on which the samples were hybridised and scanned. UHRR samples are labelled as C1-25. Replicate breast tumour clinical samples are identified with a suffix of 'a' through 'd'. The pre- and post-treatment biopsy samples are identified by a triangle to the left and right of the sample IDs, respectively. **B**: Schematic comparison of the sample pre-processing used in the generation of our Ref-8/HT-12 data, the MAQC Illumina/Affymetrix data, and the Paterson Affymetrix data.

In this study we generated two new datasets using Illumina HT-12 BeadChips (see Figure 4.1A for hybridisation plan), re-analysed the Illumina and Affymetrix MAQC datasets [118], and re-analysed a comprehensive set of technical-replicate Affymetrix arrays produced at the Paterson Institute within the University of Manchester, UK, and described previously [126]. A schematic diagram of our Illumina study design, the MAQC study design, and the Paterson study design is provided in Figure 4.1B.

1. Relative levels of technical variation in Illumina and Affymetrix microarrays

As in the first experiment using Ref-8 BeadChips, 'experiment 1', we hybridised to the new HT-12 chips an aliquot of freshly extracted UHRR that was split and frozen after amplification and labelling. On two of these arrays we also included a sample of the 'original' UHRR used in experiment 1 for consistency (see Figure 4.1A and methods). The Ref-8 (v2) and HT-12 (v3) chips share a common sample-preparation procedure, provided by Illumina, and a large number (17,542) of common probes, for which the exact same 50-nucleotide sequence is conserved, enabling a simple and direct comparison of expressions reported by these probes.

Using this combined Ref-8 and HT-12 dataset, in conjunction with the MAQC datasets and the previously published Affymetrix dataset, we report an assessment of various sources of technical variation introduced during experiment design and sample-preparation in terms of the impact on expression measurements.

2. Utility of pooled 'analysis samples' as batch-calibrators, comparison with UHRR

In the previous analysis of Ref-8 chips, we observed that the consistency of genes reported as differentially expressed was slightly improved when the UHRR control samples were included as a calibrator during the normalisation of the batch effect. We speculated that pools of clinical samples, or whatever RNA source is of direct interest to the current investigation, might perform more effectively as a batch-calibrator. This hypothesis is based upon the assumption that some genes expressed in the samples of interest will not be represented in the UHRR, but would be present in a pool of clinical samples.

We report a comparison of UHRR with pooled breast-tumour RNA in terms of the extent to which probe-level expressions are affected by batch effects and the correlation of these differences with those observed in the duplicate clinical tumour samples. The control sample in which probe-wise expressions are affected similarly to those of the clinical samples will likely perform as better calibrators during correction for these

technical batch effects.

3. Investigation of individual probe-properties in terms of a correlation with estimated levels of batch-variation

In the analysis of technical variation in the UHRR replicates there is greatly reduced scope for the introduction of 'standard' sources of noise into the reported expressions, such as that resulting from sampling, amplification efficiency, and labelling. Therefore a major remaining question is over the source of the batch variation. Many articles have reported issues with the design of microarray probes, both in terms of the quality of their mapping to the host genome [171] and in terms of their thermodynamic properties [190, 191].

In their original article describing the BeadArray platform, Kuhn *et al.* [34] provide a brief description of the custom-built probe-design pipeline, but do not provide many specifics, adding that the pipeline will "*be described in detail elsewhere (P. Rigault, in prep.)*". Unfortunately, there is no record of this article's subsequent publication. The brief details provided in [34] describe selection of appropriate genomic target regions and ranking of potential probe-sequences based on proximity to the 3' end of the transcript, melting temperature, and self-complementarity.

To that end, we assessed several hybridisation-specific probe-properties in relation to the technical variation amongst the replicate UHRR samples; including the fraction of cytosine and guanine nucleotides (GC fraction) and the number of cytosine-guanine dinucleotides (CpG) in the probe sequence. We also assessed the biological-replicate clinical samples in terms of correlation of variation with properties related to sample-prep such as the proximity of the probe to the 3' end of the target gene and the number of concurrently probed known transcripts.

4.2 Methods

4.2.1 Samples

All samples in each of the three Illumina BeadChip experiments described below, were subject to the same preparation protocol: From each extraction 100ng RNA was amplified and biotinylated using Illumina TotalPrep RNA Amplification Kit (Ambion) and quantified on a Bioanalyser 2100. 750ng cRNA per sample was hybridised to Illumina HumanRef-8 (v2) or Illumina HT-12 (v3) Expression BeadChips (Illumina, Cambridge, United Kingdom) using Whole-Genome Expression Direct Hybridisation kit (Illumina) and scanned with a BeadStation 500GX (Illumina).

Methods for the MAQC Illumina Human-6 Expression BeadChip (v1) and Affymetrix U133 Plus2.0 array hybridisations are provided in [118] and methods for generating the MCF7 and MCF10A triplicate Affymetrix U133A data can be found in the original publication [126].

All raw gene expression files and clinical annotation generated in this study are publicly available from the caBIG supported Edinburgh Clinical Research Facility Data Repository (https://catissuesuite.ecmc.ed.ac.uk/caarray/) and on request to the corresponding author.

Experiment 1

Details provided in Sabine *et al.* and Kitchen *et al.* [170, 186].

Experiment 2

Four Illumina Human HT-12 (v3) BeadChips were used in this follow-up experiment, in which the chips were processed in pairs over the course of two days (again referred to as runs). A single UHRR replicate, from a fresh preparation, was hybridised to each chip (sample IDs: 'C19' - 'C22'; Figure 4.1). For consistency, two replicates of the original UHRR from experiment 1 were retrieved from storage at -80^{o}C and added to one chip on each run (sample IDs: 'Cz1' and 'Cz2'; Figure 4.1). 34 arrays over the four chips were used to analyse eight matched primary tumour biopsies, pre- and post-treatment with an IGFR inhibitor, hybridised in duplicate over the two runs (sample IDs: '101'-'116'; Figure 4.1). Pools of pre-treatment and post-treatment tumour RNA were created from the clinical samples and each pool split into four aliquots and hybridised once to each chip. As in the first experiment, all duplication/replication of clinical, reference, and pooled samples was performed after labelling and labelled samples were stored as per the manufacturer's recommendations.

Experiment 3

Three Illumina Human HT-12 (v3) BeadChips were processed over the course of a single day, to which were hybridised more UHRR samples, in two groups of replicates. The groups correspond to different labelling protocols; the first group of samples were obtained from the same amplification and labelling (Ambion) as the UHRR samples used in experiment 2 (sample IDs: 'C23'-'C25'; Figure 4.1). The second group of samples were labelled using the NuGen amplification kit (NuGen) (sample IDs: 'Cn1'-'Cn4'; Figure 4.1). The remaining samples on these BeadChips were not analysed within this study (sample IDs: '211'-'240'; Figure 4.1).

4.2.2 Statistical Methods

Gene expression changes were compared before and after treatments and between responders and non-responders using Bioconductor [175] algorithms implemented in the statistical programming language, *R* (v.2.12.1) [176]. Illumina probe profile expression data were normalised by quantiles and corrected for batch processing effects using *ComBat* [127]. Genes differentially expressed between paired pre- and post-treatment samples were identified using *limma* (v.3.6.9) [66] and *SAM* (v.1.28) [67]. For the analysis using *limma*, genes were defined as being differentially expressed after satisfying a minimum fold-change of ± 1.5 and a maximum, Benjamini-Hochberg adjusted, p-value of 0.01. For the *SAM* analysis, the differentially expressed genes were selected at a maximum predicted false discovery rate of 5% and the same minimum fold-change of ± 1.5.

All Illumina BeadChip data were filtered, where specified, using the detection confidence reported by the *BeadStudio* software- determined for each bead based on the expressions of internal control probes, local background intensity, and the uniformity of the reported intensity of the bead. The filtering was performed prior to quantile normalisation such that probes with a detection confidence less than or equal to 95% in more than 20% of the samples were removed from further analysis. Affymetrix data were filtered such that probes reported 'absent' in more than 20% of the samples were removed and subjected to quantile normalisation.

We applied a linear additive model to expression data on the log-scale to estimate the inter- and intra-batch variance contributions. These contributions are assumed to be independent and randomly drawn from log-normal distributions. As all factors meet in unique combinations a nested, or hierarchical, variance model is individually applied for each gene. Models of this kind are formally defined in [192] and have previously been used in the context of gene-expression experiment design [144, 193]. Variance estimates in all analyses described herein were performed using an REML procedure implemented in the *nlme* package in *R* [194, 195, 196]. In all mixed models the biological variables such as different cell-lines in the Paterson dataset and the UHRR/UBRR dilutions in the MAQC dataset were treated as fixed effects and all downstream sample-processing levels treated as random effects.

Probes were re-mapped to the human genome (NCBI build 37) using *Bowtie* [197] (v. 0.12.7) allowing for no mismatched bases in the alignment. Alignment annotation, for example the position of the probe within the host gene, was provided by in-house software and the NCBI RefSeq annotation database [198]. The 309 probes common to both the Ref-8 and HT-12 chips that aligned the genome but fell within intergenic regions (or those that, for whatever reason, could not be annotated using the

RefSeq database) were considered good and retained, along with the 15,448 annotated intragenic probes, for further analyses.

4.3 Results

4.3.1 UHRR analysis using our Ref-8 and HT-12 data

To perform reliable correlation and variance analyses simultaneously on both Illumina chip-types, we identified and retained only the 15,757 probes with exactly conserved sequences between Ref-8 and HT-12 that also mapped uniquely to the genome (see methods). All data were detection-filtered and normalised together (8,948 of the conserved probes passed the filter, see methods) before assessment by pairwise Pearson correlation (Fig 4.2A and Fig 4.2B). Results of pairwise Spearman rank correlations following separate filtering and normalisation of the Ref-8/HT-12 chips were very similar (data not shown). The globally filtered/normalised data were used in all subsequent UHRR analyses for consistency with our previous results and for a more reliable interpretation of nested variance analyses.

It is clear from the heatmaps in Figure 4.2 that the majority of the variation between replicate UHRR samples exists due to the three separate experiments. However, somewhat surprisingly, the poorest correlations exist between the two experiments involving the HT-12 chips (85.6% DF+QN between runs 6/7 and run 8) as opposed to (90.4% between runs 1-5 and run 8). As previously reported, standard normalisation techniques based on array-wide intensity distributions do little to improve probe-wise correlations between replicate samples [186]. There is also an apparent band of poorer correlation corresponding to a single HT-12 chip in run 7 (chip 22).

For consistency with the control samples hybridised to the Ref-8 chips, two replicates of the original UHRR from experiment 1 were retrieved from storage and added to one chip on each run in experiment 2. However, there was only a marginal improvement in the correlation of these old UHRR samples compared to the freshly amplified and labeled replicates (C1-18 vs. Cz1:Cz2 = 92.8% compared to C1-18 vs. C19:22 = 91.2%) although, again, the sample on chip 22 was more poorly correlated with the UHRR samples on the Ref-8 chips than that hybridised to chip 20.

All UHRR data were subjected to batch-correction using *ComBat* and correction at two levels was assessed: experiment-wise and run-wise correction. Following *ComBat* correction by experiment, all pairwise correlations increased to approximately the same level observed in the quantile-normalised Ref-8 data obtained from the first experiment (Fig 4.2C). However, we found previously that despite high correlation between UHRR replicates the consistency between lists of statistically significant differentially expressed

probes from duplicate sets of samples was poor without specific batch-correction [186]. Following *ComBat* correction by run, correlations in these new UHRR samples approach those observed in the original, *ComBat* batch-corrected UHRR samples from the first experiment (Fig 4.2D).

In addition to correlating UHRR expressions between sample-pairs we performed variance estimates, for each of the 15,757 probes, at the inter-experiment, inter-run, and inter-chip using a nested analysis of variance described in methods (Fig 4.3, main panel). As expected, and in agreement with correlations in Figure 4.2, the experiment was the parameter with the greatest source of measurement noise, accounting for an average 61.7% of the total variation in reported expressions. This is likely due to the use of a fresh round of amplification and labelling performed on the new UHRR samples. Despite an overall reduction in the technical variation between probes, the fraction of the total variance contributed by the each level was unaffected by quantile normalisation. Again, it is clear that batch correction greatly reduces the technical variation due to experiment and run. The slight effect due to the within-batch variance moderation can be seen in the reduction in inter-chip SD after either run of *ComBat*. The high inter-chip variation, compared to inter-run variation, appears to be driven by the new samples run on the HT-12 chips as this was not observed in the standalone analyses of the Ref-8 data in experiment 1.

A similar variance analysis was performed, for comparison, using two of the MAQC array datasets. The MAQC Illumina Human-6 Expression BeadChip (v1) dataset, with expressions for 47,293 probes, was subjected to the same detection-filtering criteria as our Ref-8 and HT-12 chips and the expressions of 21,896 surviving probes were quantile normalised prior to variance analyses. The design of the MAQC experiment only allowed for variance components to be estimated for the inter-laboratory, inter-chip, and inter-array levels. However, compared to the Ref-8/HT-12 inter-experiment standard deviations, the MAQC inter-laboratory standard deviations were much smaller; less than half as much on average (Fig 4.3, right panel). This is likely to be a result of our use of different array-versions and widely different dates on which the arrays were processed (about 2 years). However it is noteworthy that the majority of the variance in the MAQC dataset was attributable to the intra-chip (inter-array) level and this is the only dataset, for which variance analyses were performed, in which this phenomenon holds.

The same variance analysis using the MAQC Affymetrix U133 Plus2.0 dataset (using 23,053 probes reported as 'present' across at least 80% of the samples) revealed a more familiar variance structure in which the majority of the variation is directly attributable to systematic differences between the different laboratories performing the experiments (Figure 4.4, left panel). The estimated standard deviations at the

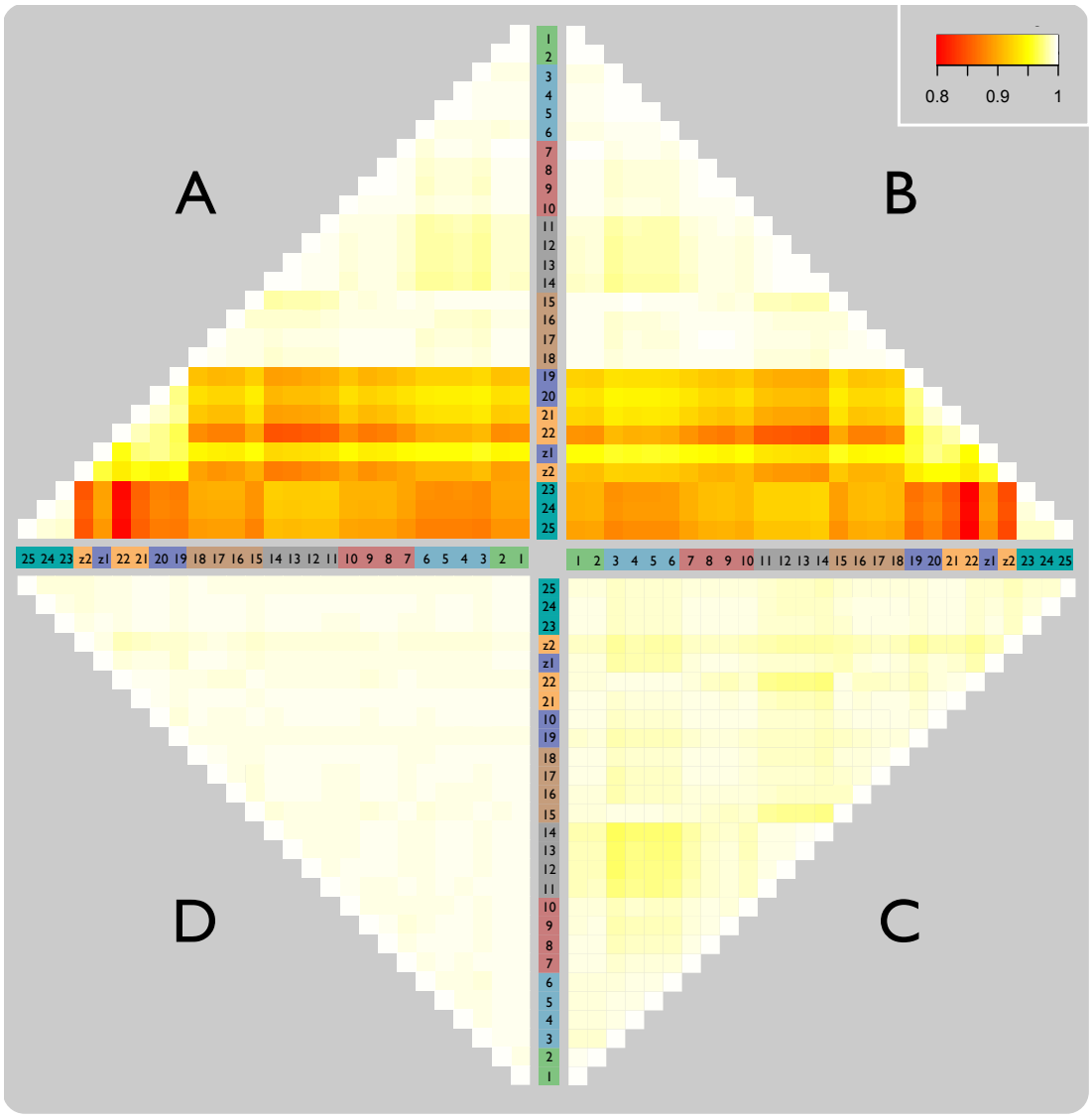


Figure 4.2: Heatmap of Pearson correlations between replicate pairs of UHRR samples highlights the inter-experiment, inter-run, and inter-chip differences; particularly at the inter-experiment level. Red cells correspond to ∼80% correlation and white to 100% correlation. Batches and sample numbers are consistent with the colouring and labelling in Figure 4.1A. **A**: detection filtered (DF); **B**: DF & quantile normalised (QN); **C**: DF & QN & *ComBat*(by experiment); **D**: DF & QN & *ComBat*(by run).

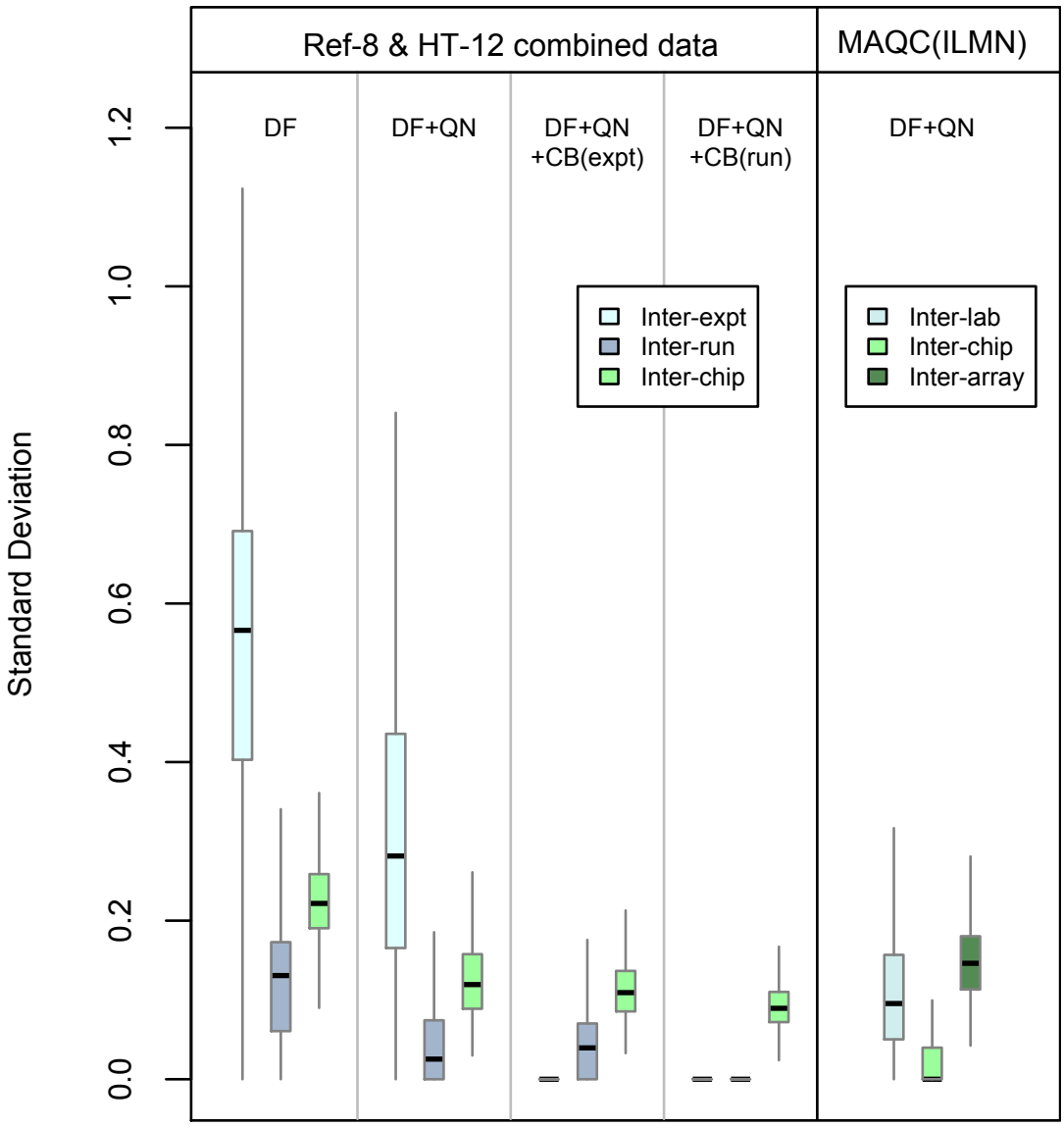


Figure 4.3: Comparison of variance components in our Ref-8/HT-12 data to the MAQC dataset. **Ref-8 & HT-12 combined data**: Probe-wise estimates of standard deviations (SD) corresponding to the inter-experiment (light blue), inter-run (dark blue), and inter-chip (green) technical variance in our UHRR data. The effect on these standard deviations following detection-filtering (DF), quantile-normalisation (QN), *ComBat* batch-correction by experiment (CB(expt)), and *ComBat* batch-correction by run (CB(run)) are shown.
MAQC(ILMN): Probe-wise SD-estimates corresponding to inter-laboratory (pale blue), inter-chip (green), and inter-array (dark green) technical variances in detection-filtered and quantile normalised UHRR/UBRR expressions from the MAQC Illumina dataset.

inter-laboratory level in these data are similar to the inter-experiment level in the Ref-8/HT-12 data, as are the distributions of the coefficients of variation (standard deviation normalised by mean expression; data not shown).

Finally, variance analysis of a previously described Affymetrix dataset of multiple replicates of the MCF7 and MCF10A breast cell lines [126] provided a far more detailed breakdown of the estimated biological, as well as technical, error introduced at the various levels of sample-preparation prior to an array experiment (Figure 4.4, right panel). The majority of the probesets show far greater inter-sample variability than between the two cell-lines. Not obvious from this plot however, are the 4,302 probesets for which the inter-cell-line is greater than the inter-sample standard deviation, or the 659 probesets for which it is greater than the sum of all lower levels, potentially indicating a fairly large amount of differential expression between the cell-lines. Inter-sample variability, averaged over all probesets, contributed 40.6% to the total standard deviation; compared to 13.9% due to amplification/labelling, 9.9% & 10.8% for inter array and inter-scanner, and 15.2% at the within-scanner/residual level. This error profile is not at all dissimilar to error estimates obtained from the various stages of sample preparation prior to a qPCR experiment in which, when using solid tissue, the sampling step is by far the most variable while the noise introduced during reverse-transcription is generally low, but is sometimes larger than sampling [193].

4.3.2 Inter-batch calibrators: Comparing UHRR with pools of tumour sample RNA

In addition to the repeat hybridisation of UHRR replicates, two pooled sample controls of clinical breast-tumour RNA were run on each of the four BeadChips in experiment 2 (samples P1 through P8; Figure 4.1A). One pool was created from a mix of all seven pre-treatment samples run in this experiment and the second pool was created using all seven post-treatment samples (see methods).

Results of pairwise Pearson-correlations between these pooled samples identified a large difference between chip 22 and the other three chips used in this experiment (Fig 4.5A & 4.5B), consistent with that seen with the UHRR. No obvious differences were observed in the correlations between the different pools composed of either pre- or post-treatment RNA. In general, the correlations appeared similar to those observed between replicate UHRR samples illustrated in Figure 4.2. As noted before, quantile normalisation does little to remedy the poor correlation between chip 22 and the other chips, but this is remedied by *ComBat* by treating the batches as either as runs (Fig 4.5C) or, slightly better, as separate chips (Fig 4.5D).

Compared to the UHRR samples in experiment 1 (Fig 4.6A), variance components

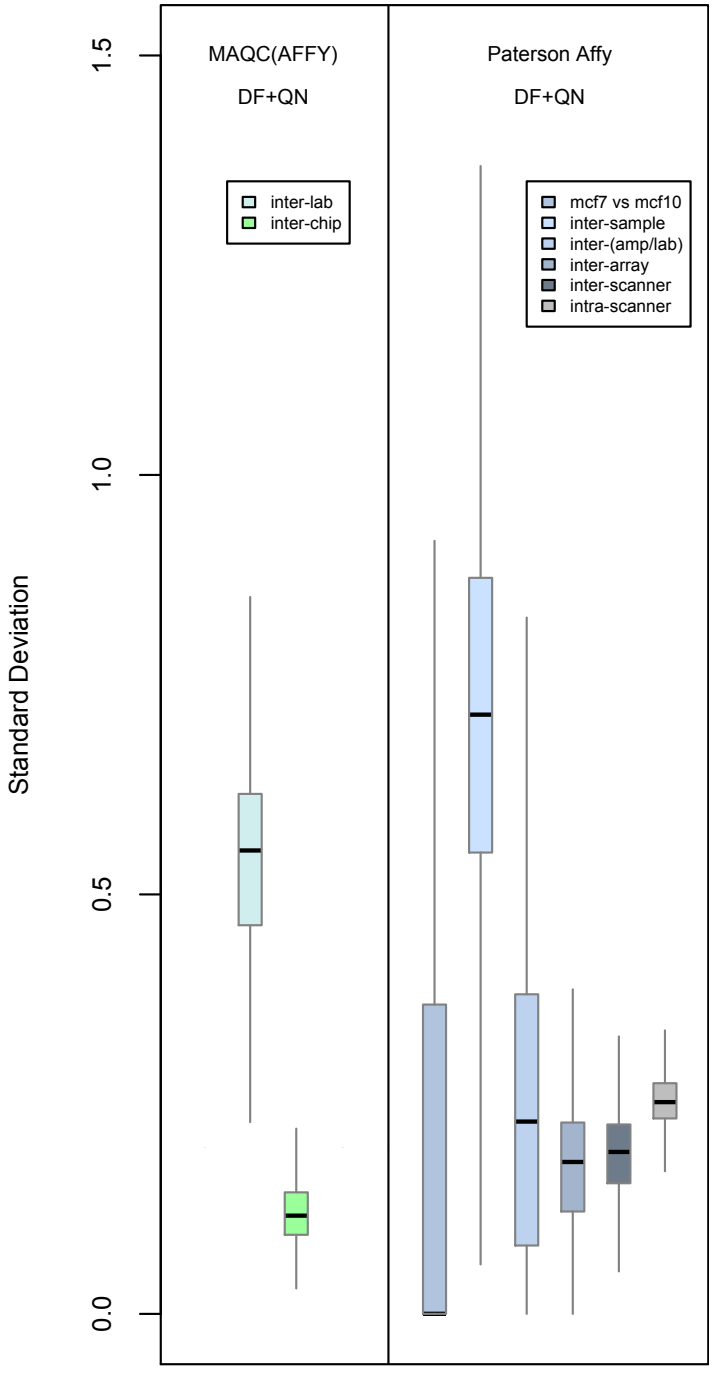


Figure 4.4: Comparison of MAQC and Paterson Affymetrix variance components. **MAQC(AFFY)**: Probe-wise SD-estimates corresponding to inter-laboratory (pale blue) and inter-chip (green) technical variances in detection-filtered and quantile normalised UHRR/UBRR expressions from the MAQC Affymetrix dataset. **Paterson Affy**: Probe-wise SD-estimates corresponding to several levels of technical variance (see figure key) in detection-filtered and quantile normalised MCF7/MCF10 expressions from the Paterson Affymetrix dataset.

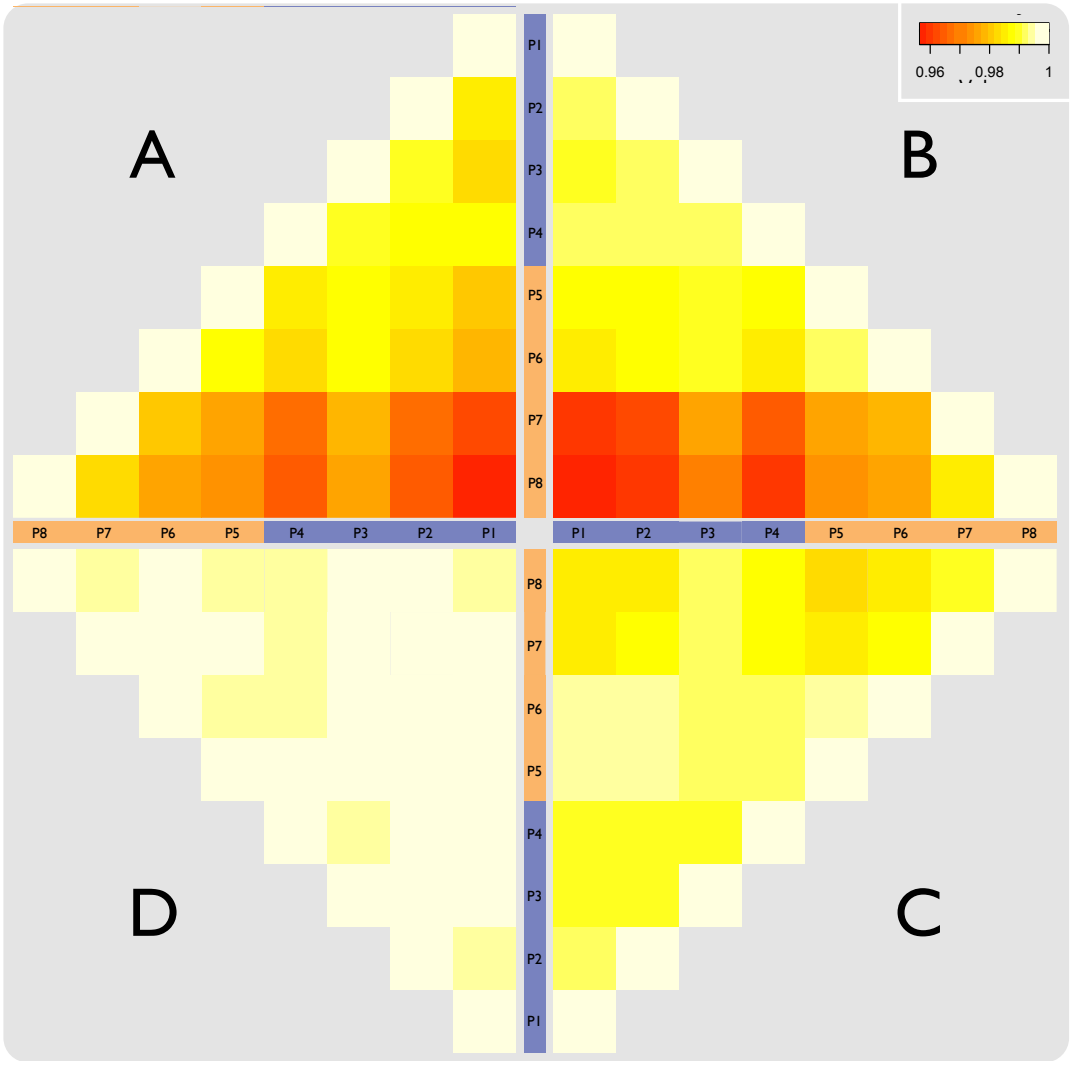


Figure 4.5: Heatmap of Pearson correlations between replicate pairs of pooled-tumour control samples highlights the inter-run and inter-chip variation; particularly at the inter-chip level. Red cells correspond to ∼96% correlation and white to 100% correlation. Batches and sample numbers are consistent with the colouring and labelling in Figure 4.1A. **A**: detection filtered (DF); **B**: DF & quantile normalised (QN); **C**: DF & QN & *ComBat*(by run); **D**: DF & QN & *ComBat*(by chip).

estimated by nested-anova between the new UHRR (Fig 4.6B) and pooled tumour samples (Fig 4.6C) in experiment 2 both show a large inter-chip/intra-batch effect. When quantile normalisation is applied, the inter-run variation is resolved, albeit at a low level. Overall estimates of standard deviations in Fig 4.6B are larger than 4.6C which is, in turn, larger than Fig 4.6A, however this increase is likely due to the smaller number of replicates in the later experiment making it harder to accurately estimate the probe-wide error-levels. The larger magnitude of variation explains the increased inter-chip component in Figure 4.3 and it is clear from both the correlations and the variance estimates of both the UHRR and tumour pools that the variance in this second experiment is driven by a particular chip (chip 22), rather than a particular run as was the case in the first experiment.

Differential expression analyses were performed on the pre- and post-treatment tumour duplicate samples using *limma* and *SAM* (see methods). Each run within experiment 2 was treated as a standalone duplicate sub-experiment, and as such data were preprocessed, normalised, batch-corrected by chip, and analysed for differential expression completely independently from each other. Run 6 corresponded to sub-experiment 1 and run 7 to sub-experiment 2. Data from each run were independently quantile normalised, batch-corrected by chip using only the tumour samples, batch-corrected by chip using the tumour samples and the UHRR replicates as controls, and batch-corrected by chip using the tumour samples and the tumour-pool replicates as controls.

Unfortunately only 5 probes that were found significant in these tests also survived multiple-testing correction at $q<0.05$, and the same 5 probes appear in every results list regardless of normalisation or batch-correction. The number of probes satisfying $p<0.05$ vary between the two sub-experiments; more probes were identified as significant on the analysis of run 6 compared to run 7. However the fraction of significant probes consistently identified in both sub-experiments is $\approx 20\%$ suggesting high number of false positives making this inappropriate for further analyses. Analysis using *SAM* produced the same result, with the same 5 probes consistently differentially expressed in both replicate groups, independent of normalisation/batch-correction (data not shown).

Despite the apparent low level of biological variation between the pre- and post-treatment samples and small number of significantly differentially expressed genes, both the UHRR and the Pooled tumour accurately reflected the fold-changes between duplicate tumour samples across the two runs (Fig 4.7). The plot highlights the increased correlation of both the UHRR and the two pool controls with the individual tumour duplicates when one of the duplicates was present on the seemingly outlying chip 22 (second and fourth panels). In addition, the pools outperformed UHRR in terms of modelling the specific run-induced difference in expression, measured in terms

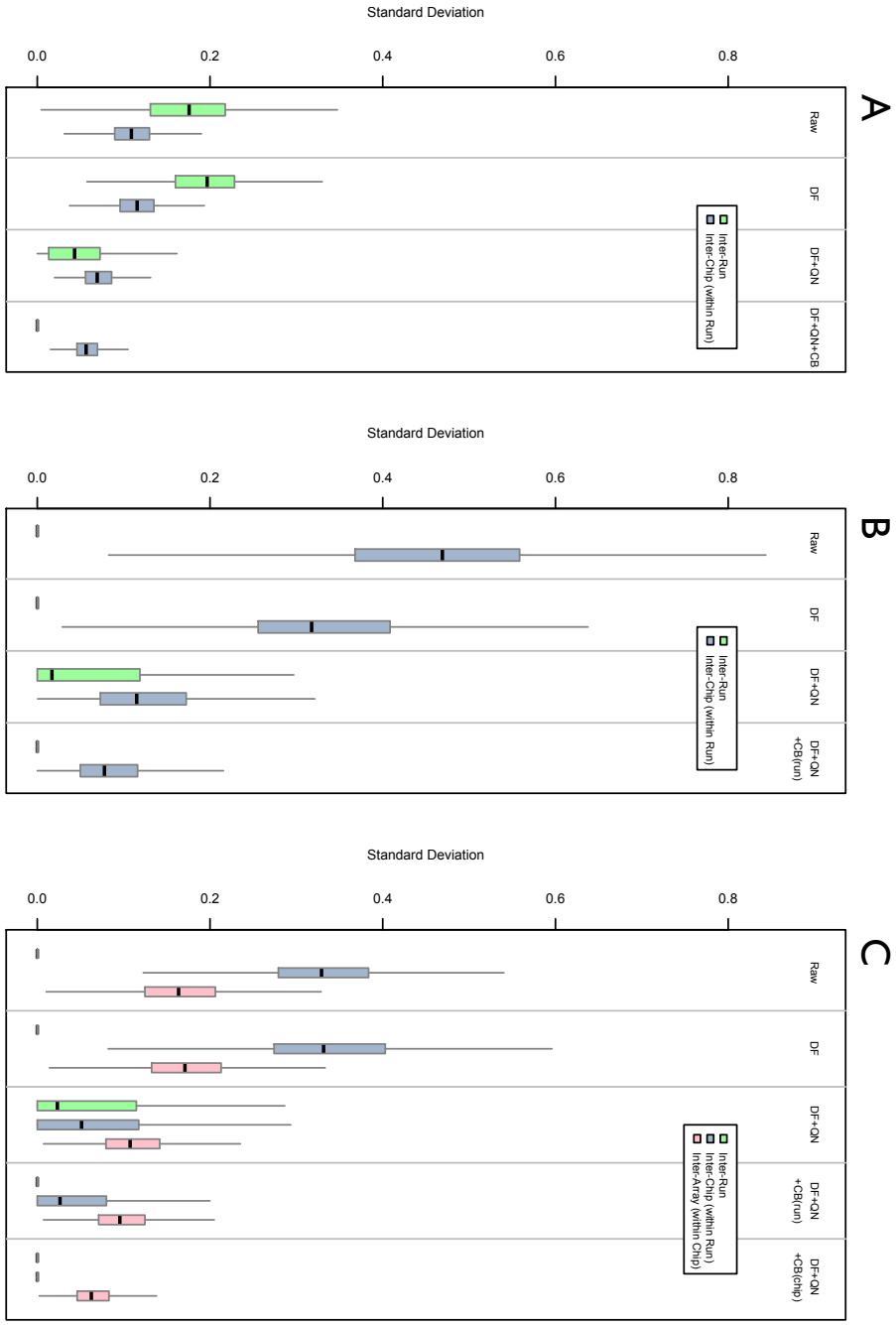


Figure 4.6: Variance estimates at various levels in replicate samples hybridised to the Ref-8 and HT-12 chips used in experiments 1 and 2. Also shows the effect of various normalisation procedures on these variance estimates; as before such procedures include detection-filtered (DF), quantile normalised (QN), *ComBat* corrected by run (CB(run)), and *ComBat* corrected by BeadChip (CB(chip)). **A**: Inter-run and inter-chip variance estimates using UHRR replicates in experiment 1 (previously reported in [186]). **B**: Inter-run and inter-chip variance estimates using UHRR replicates in experiment 2. **C**: Inter-run, inter-chip, and inter-array variance estimates using pooled-tumour replicates in experiment 2.

of correlation with the same differences as observed between tumour duplicates across the chips. The pre- and post-tumour pools were found to increase correlation, on average, by 3.9% and 4.3%, respectively, compared to UHRR; one-sample t-tests based on the difference in correlations between these pools and UHRR were both highly significant ($p=5.9 * 10^{-9}$ and $p=3.4 * 10^{-9}$, respectively).

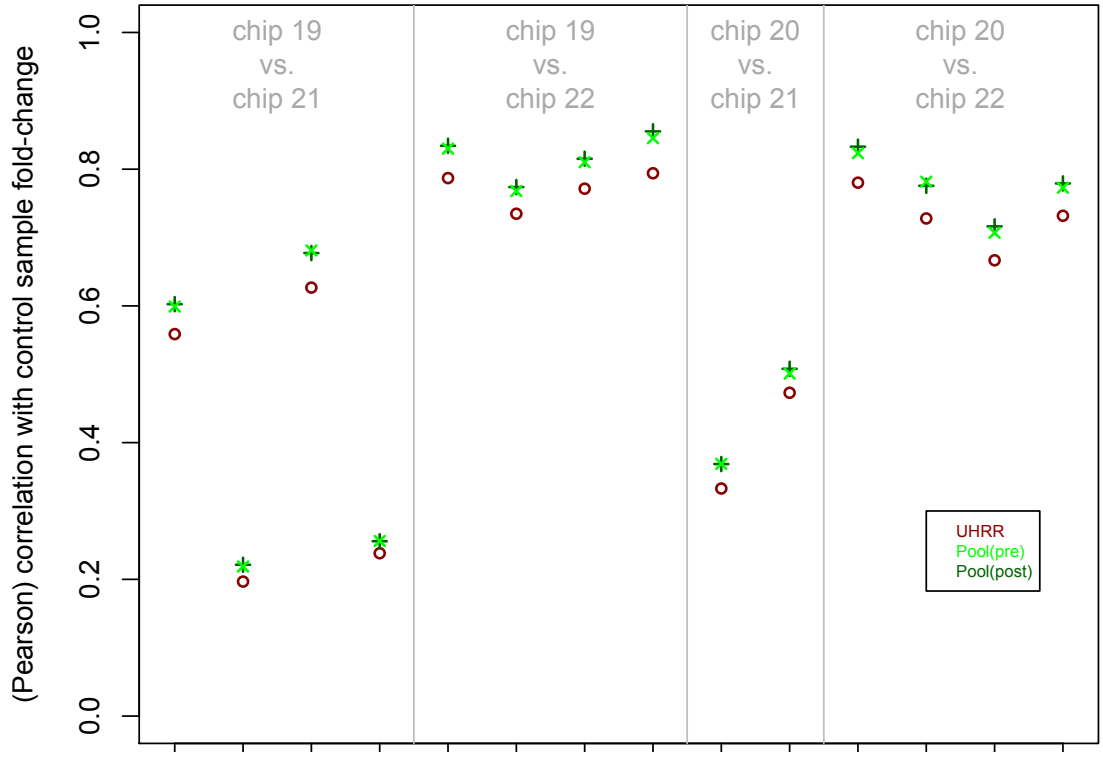


Figure 4.7: Correlation of expression change as a result of inter-run and inter-chip technical variation between UHRR and pooled controls with tumour duplicates. Tumour duplicates (individually plotted) are arranged on the x-axis to be close to others processed on the same BeadChip. UHRR and both types of pooled-control (comprised of pre- and post-treatment tumour RNA, respectively) are correlated more strongly with individual tumour duplicates in which one 'half' of the duplicate was processed on the outlying chip 22. Pooled-controls also consistently score slightly higher correlation than UHRR.

We performed an exhaustive comparison of the fold-change in expression due to the different runs between the replicate pools of pre-treatment tumour RNA and each of the fold-changes between duplicate tumour samples (Figure 4.8). These plots each show the magnitude of the change in expression between technical replicates introduced by the runs and show how well correlated such changes are between the Pool/tumour-duplicates. In the figure, the pre-treatment tumour duplicate are coloured blue and

post-treatment are green; probes that are differentially regulated, up or down, at least two-fold due to the different runs are highlighted in each plot (red points). Note that most of the samples with a duplicate on chip 22 are subject to a greater magnitude of variation than samples with a duplicate on chip 21. In many cases there is a reasonably strong agreement between absolute expression differences observed in the pool and the individual tumour duplicates, i.e. follow a line of unit gradient passing through the origin. However such agreement is not necessary in order for the effect to be removed by *ComBat*, or alternative methods, as the absolute magnitude of the effect is explicitly normalised during the correction procedure. Instead, it is desirable to have a strong correlation between the control sample and the test samples for probes subject to large differences in expression arising from technical experiment noise; it is clear that the pooled-tumour controls (of both pre- and post-treatment samples) are consistently more highly correlated than the UHRR (Figure 4.7).

4.3.3 Properties of the probes with respect to batch variation

In order to assess the technical systematic-variation in terms of properties of the probes themselves and variables implicit in their mapping to the reference genome, we mapped all probe sequences on the Ref-8 and HT-12 arrays to the human reference genome (see methods). Several straightforward descriptive statistics were elected to serve as measures of probe- and mapping-specific properties that could conceivably influence probe expression. These included compositional properties such as the guanine and cytosine nucleotide content (GC) and cytosine-guanine dinucleotide (CpG) content of the probes, as well as mapping properties such as the position of the probe as a fraction of the total length of the target gene, the number of (known) transcripts consecutively probed, and the average number of exons within the probed gene; with the number of known transcripts and number of exons acting as proxies for gene complexity and size, respectively.

We used the analysis of run-induced fold change between duplicate tumour samples in experiment 2 as a platform to explore potential correlations between probe GC content on the observed variations. These samples were split into duplicates after labelling so each half of a duplicate pair was subjected to the same extraction/purification, the same amplification, and the same labelling; therefore any variation between them must be a result of noise introduced during the hybridisation and scanning of the arrays and therefore more likely due to composition of the probes, rather than their mapping properties.

GC content is well known to be a simple and important factor in the design of primers and probes due to its positive correlation with hybridisation stability and

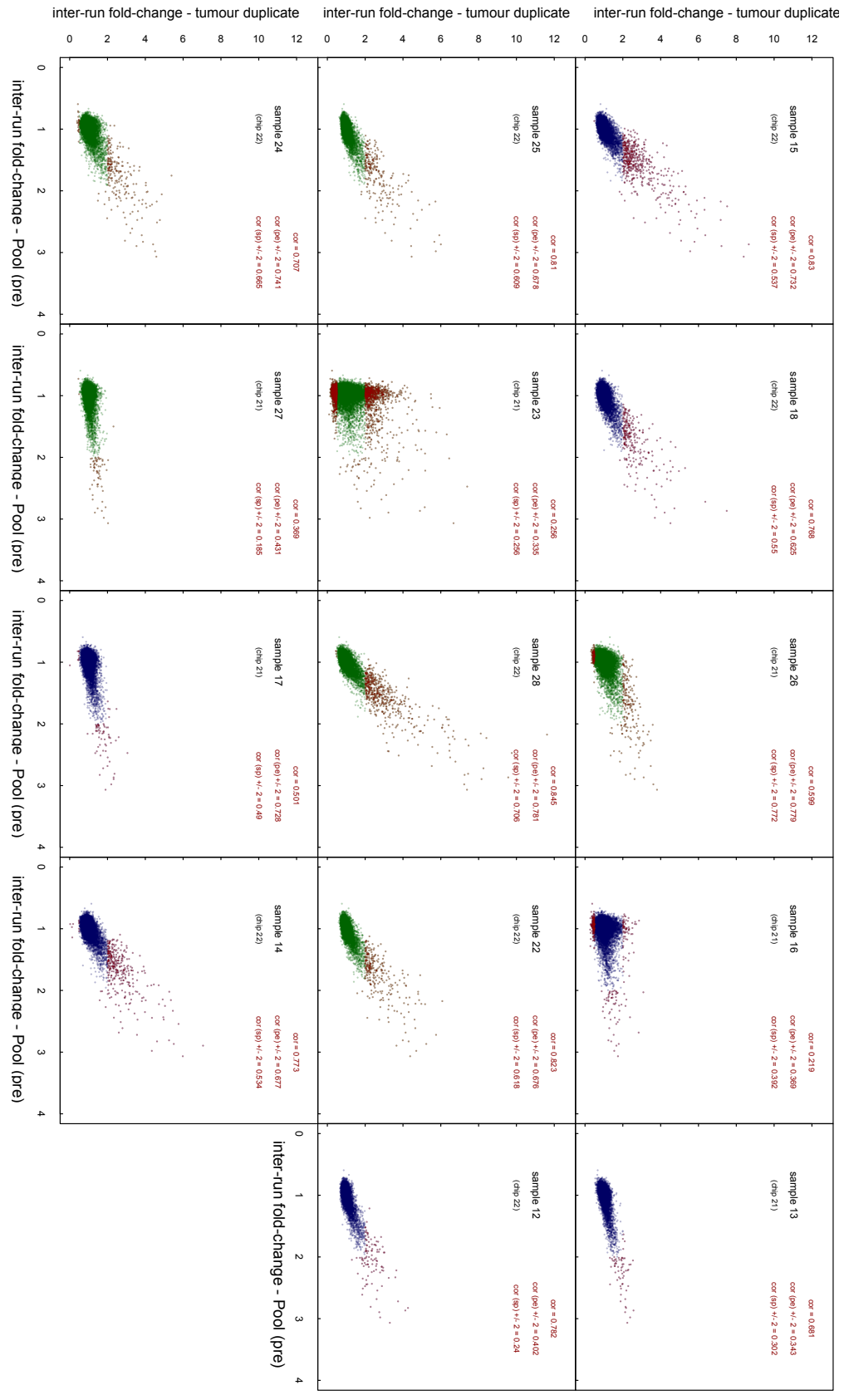


Figure 4.8: Scatter plots of pool vs. tumour fold-change.

melting temperature [190, 191, 199]; increased read density in regions of high GC content from second-generation Solexa sequencing data has also been reported [200]. In terms of the distribution of all probe GC fractions over an array, one would expect, if no specific probe-design decision had been made taking this into consideration, a normal distribution about 0.5, i.e. we expect, on average, 50% bases to be either a guanine or a cytosine. The distribution of GC content over our Illumina HT-12 probes (Figure 4.9A) clearly shows that this is not the case and Illumina have specifically designed the probes with a bias in favour of greater-than-random GC content.

In Figure 4.9B the GC-fraction distribution of the 12,042 probes used in the analysis of run-induced fold change is plotted as a Gaussian-smoothed probability density. Also plotted is the distribution of the subset of these probes found to be more than 2-fold up- or down-regulated due to the batch effect in any of the 14 duplicate sample pairs. These two distributions are very similar, suggesting that a large number of these technical effects are random fluctuations independent of GC content. However the distribution of any probe found to be more than 2-fold up- or down-regulated due to the batch effect in at least 6 of the duplicate sample pairs is clearly biased towards lower GC fractions. Results of a chi-square test based on probes with less than 50% GC content were highly significant (χ^2, pVal $< 2.2 * 10^{-16}$). This suggests that while the majority of probes may be randomly affected by the batch-effect, a core number (207) of probes with lower than average GC content are consistently affected in our experiments by technical variation due to batch-processing of the arrays.

Taking this further, we correlated probe GC-fraction with the inter-experiment, inter-run, and inter-chip error estimates from the UHRR replicates, again using only the conserved probes on both the Ref-8 and HT-12 arrays. We found a highly significant trend in which probes with lower GC fraction exhibited higher overall technical variation. When the technical error estimate is plotted against GC content a trend towards higher standard deviations (SD) at lower GC fractions is obvious at both the inter-experiment and inter-run levels (Fig 4.10). Chi-square analyses showed this low-GC/high-SD effect to be highly significant at both the inter-experiment and inter-run level. Probes were defined to be low-GC if their GC-content is below 0.55 as approximately 50% of the conserved probes plotted had GC-content below this level. The choice of standard deviation cutoff defining high-SD was somewhat arbitrary and depended on the general distribution of the scatter at each level. Both the GC-cutoff and the SD cutoffs used are highlighted in Figure 4.10. At the inter-experiment level, Chi-square analysis showed a significant enrichment for low-GC, high-SD probes compared to high-GC, high-SD probes (664 vs. 469; χ^2 pVal $= 5.3 * 10^{-8}$). An even more significant effect was observed at the inter-run level where low-GC, high-SD probes outnumbered high-GC, high-SD probes more than 8 to 1 (350 vs. 41; χ^2 pVal

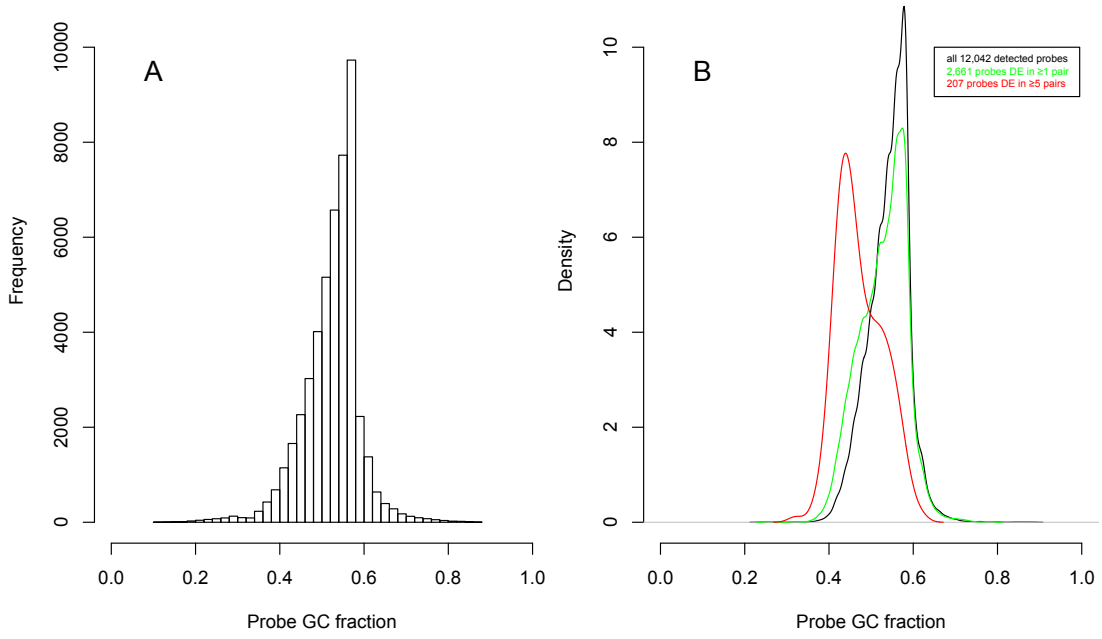


Figure 4.9: **A**: Histogram of GC content over our Illumina HT-12 probes. **B**: Gaussian-smoothed probability density distribution (black, N=48,803). Union of probes more than 2-fold up- or down-regulated due to the batch effect in any of the 14 duplicate tumour-sample pairs from Figure 4.8 (green line, N=2,661). Union of probes more than 2-fold up- or down-regulated due to the batch effect in more than 5 of the duplicate tumour pairs (red line, N=207)

$< 2.2 * 10^{-16}$). There was no significant enrichment at the inter-chip level.

A similar observation regarding probe GC content and expression consistency was recently reported in a comparison of RNA preservation protocols, using matched samples, in terms of the effect on results of downstream expression analyses [201]. To further extend these analyses, we also compared probe GC content with our variance estimates from the MAQC Illumina and Affymetrix datasets (Figure 4.3 and 4.4, respectively). These MAQC datasets have a similar overall distribution of probe-GC content, skewed in favour of higher GC fraction (data not shown), again in which approximately 50% of probes have a GC content below 0.55. Again, GC-fraction was plotted against standard deviation estimated at each of the inter-laboratory, inter-chip, and inter-array levels (Figure 4.11). We performed chi-square analyses, again using the GC-fraction cutoff of 0.55 and somewhat arbitrary standard deviation cutoffs illustrated in Figure 4.11. At the inter-laboratory level, almost twice as many low-GC probes had an estimated standard deviation exceeding the chosen boundary as high-GC probes

(173 vs. 100; χ^2 pVal $= 9.8 * 10^{-8}$). A similar, but less significant result was observed at the inter-chip level (χ^2 pVal $= 2.1 * 10^{-4}$).

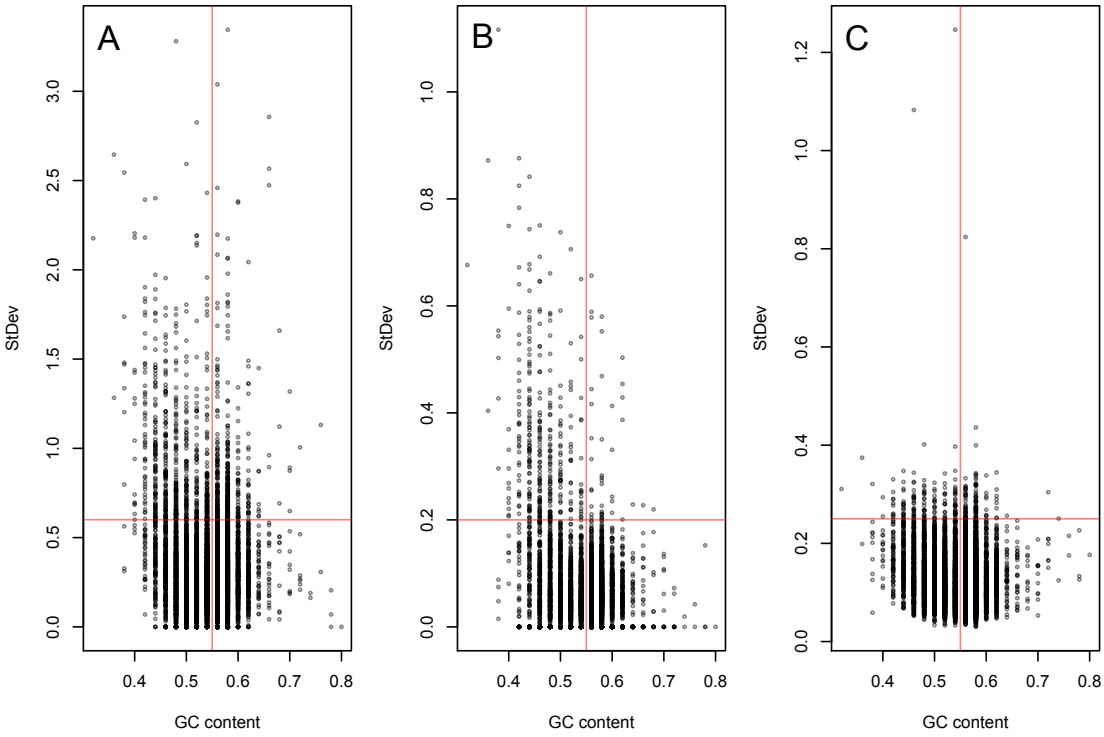


Figure 4.10: Plots of probe CG-fraction against probe standard deviation estimated at the inter-experiment (**A**), inter-run (**B**), and inter-chip (**C**) levels in our combined Illumina Ref-8/HT-12 dataset. Red lines denote the cutoffs used in chi-squared analysis at each level

In an attempt to detect whether this low-GC-high-SD effect is resolvable between biological, as well as technical, replicates we chose a subset of pre-and post-treatment samples from the Ref-8 dataset in experiment 1. We were careful to avoid confounding this analysis with inter-run variation in these data and selected, from detection-filtered and quantile normalised data, all pre-treatment biological replicates from run 3 and all non-duplicated post-treatment biological replicates from run 5. Despite a slightly greater number of probes with low-GC-high-SD in both the pre- and post-treatment sample-sets, there were no significant effects reported by chi-square analysis (data not shown).

In an attempt to detect any effects of probe-mapping properties on in our array data, we used the same set of pre-treatment biological replicates from run 3 and non-duplicated post-treatment biological replicates from run 5. These biologically distinct samples, from different experimental subjects, underwent different extractions

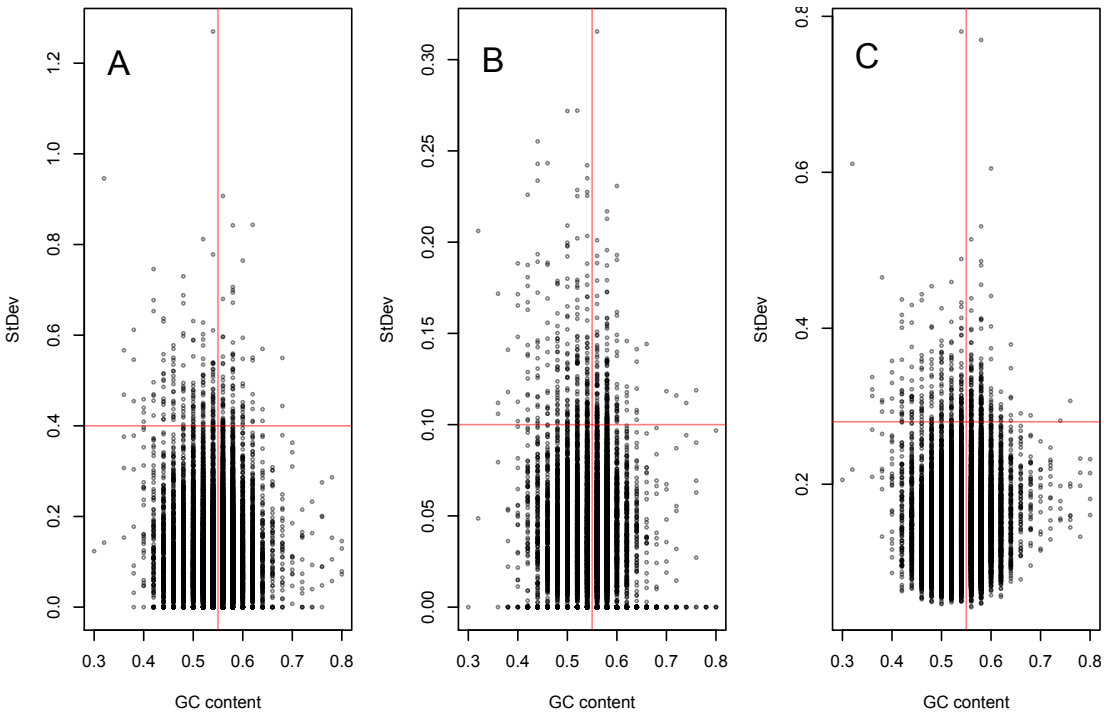


Figure 4.11: Plots of probe CG-fraction against probe standard deviation estimated at the inter-laboratory (**A**), inter-chip (**B**), and inter-array (**C**) levels in the MAQC Illumina dataset. Red lines denote the cutoffs used in chi-squared analysis at each level.

and therefore any biologically driven, mapping-specific artefacts could potentially be resolved. For each probe, the standard deviation between the pre-treatment samples and, separately, the post-treatment samples were assessed against probe position in the target gene, number of concurrently probed transcripts, and the number of exons in the target gene. Illustrated in Figure 4.12 is the distribution of probe-location as a fraction of gene length among the Ref-8 probes. After adjusting for the strand of the probe target, no significant enrichment, again assessed by chi-square, was observed for the standard deviations of probes proximal to either the 3' or 5' positions, probably due to high biological variability (data not shown). Biologically equivalent replicates, for example several independent samples from the same tissue, might provide a means to potentially resolve such subtle probe-mapping artefacts, however such replicates are not present in our dataset.

To test whether the probe-mapping may correlate with replicate cRNA-synthesis, a similar analysis was performed using the MAQC Illumina dataset. These probes have a distribution similar to the Ref-8 arrays shown in Figure 4.12 and the analyses again revealed no significant enrichment for estimated standard deviation due to probe

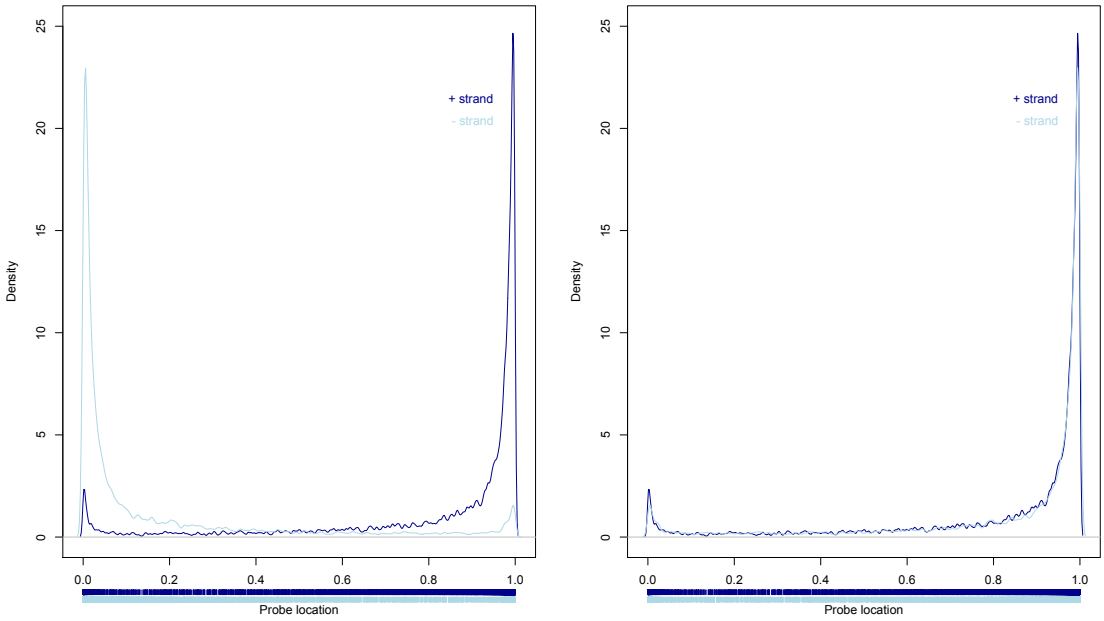


Figure 4.12: Distribution of MAQC Illumina probes as a fraction of target gene length. Light blue points are used for probes that mapped to the antisense DNA strand and dark blue for the sense strand. Left plot represents the fraction of target gene length in terms of 3' and 5' coordinates on each strand while the right plot is 'normalised' for the anti-sense strand and plots the fraction in terms of absolute position along the DNA molecule.

position (Figure 4.3.3). It is therefore likely that such probe-mapping effects are only able to be resolved when probes target different regions of the transcript, rather than probes targeting the same region being affected by biological or technical variability.

4.4 Discussion

Microarrays represent a powerful means of rapidly assessing genome-wide expression patterns for clinical applications. Unfortunately confounding technical variation and systematic error in array technologies presents a major obstacle to their adoption for clinical diagnostics in humans. Building on a previous investigation of technical variation between replicate RNA samples from breast tumour biopsies, this extended study used both Illumina and Affymetrix arrays to explore the reliability of reported expressions across a variety of experiment designs.

Using a large set of conserved, reliably-detected probes on Illumina Ref-8 and HT-12 BeadChips we found that the correlation between replicate UHRR hybridisations in the combined Ref-8/HT-12 dataset (experiments 1 through 3) were consistently poorer

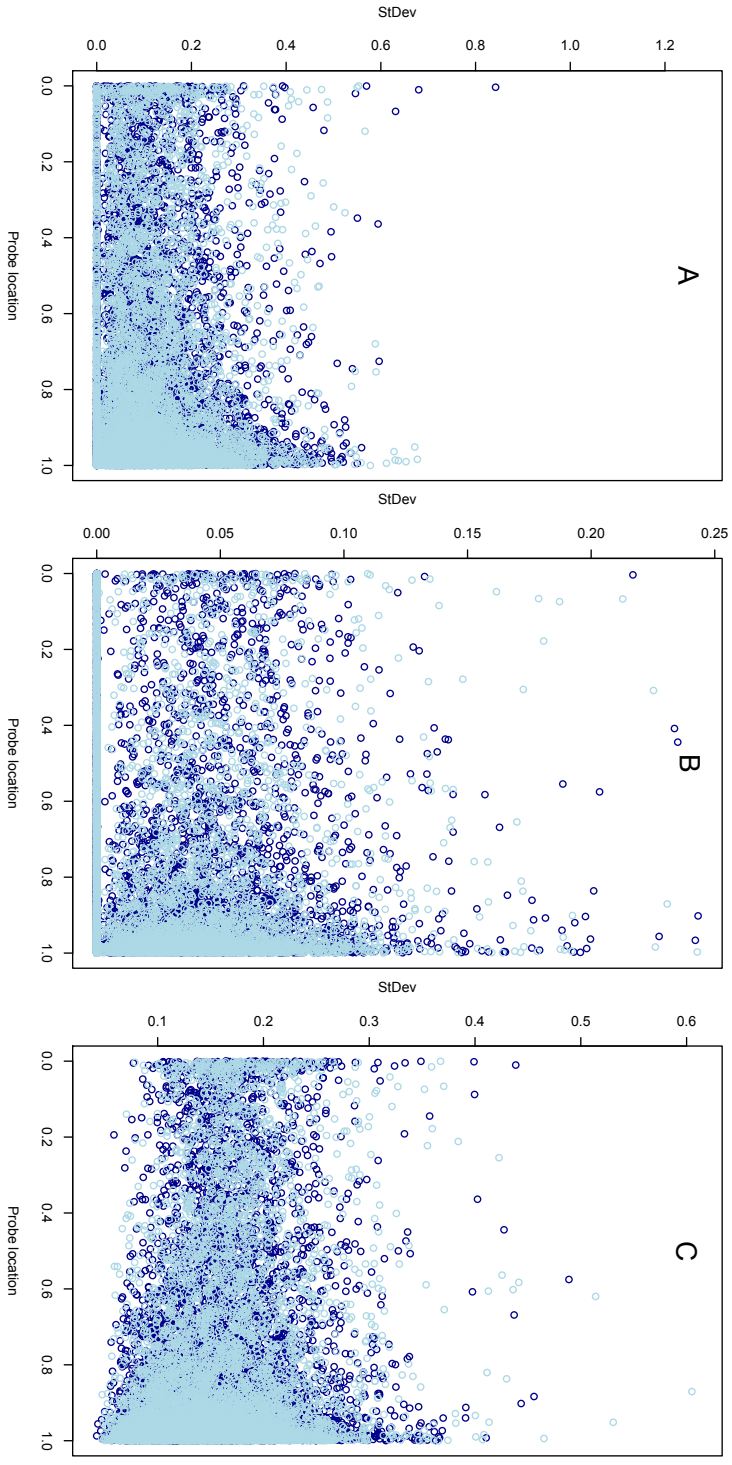


Figure 4.13: Plots of probe position against probe standard deviation estimated at the inter-laboratory (**A**), inter-chip (**B**), and inter-array (**C**) levels in the MAQC Illumina dataset. Light and dark blue points again identify probes that mapped to the antisense and sense strands, respectively.

than correlations previously reported using only the Ref-8 data from experiment 1. Interestingly, we found that UHRR samples from the original experiment, which were stored at -80^{o}C for approximately two years, hybridised to two arrays on the new HT-12 chips correlated better with the original Ref-8 samples than did freshly prepared UHRR replicates. This suggests that even long periods of frozen storage and additional freeze-thaw cycles introduce less noise into experimental measurements than that inherent in creating a new preparation of labeled cRNA, even from the same RNA source.

As in our first experiment, quantile normalisation did little to improve correlation between the UHRR replicates across the Ref-8/HT-12 dataset. However specific batch-correction using *ComBat* once again greatly improved the correlations and is a valuable tool for removing systematic error introduced between experiments and/or processing runs. Variance analyses using these combined data revealed the inter-experiment level as by far the most variable source of confounding noise, however this was confounded with different RNA preparations due to a shortage of starting material. The inter-chip variation in the new HT-12 datasets was almost double what it was in the Ref-8 dataset and due to this increase in inter-chip variation and high-levels of inter-experiment variation, the inter-run variation in the combined dataset was largely obscured. However, as we have previously seen, inter-experiment and inter-run variances were largely eliminated following *ComBat* corrections.

Variance estimates using the MAQC (Illumina) dataset were similar in magnitude to the variances obtained from just the Ref-8 data in experiment 1. This suggests an excellent level of reliability between the three laboratories that performed these experiments and also that the amplification and labelling of sample RNA for Illumina analysis introduces very little noise compared to replicate RNA extractions. In contrast, the MAQC Affymetrix dataset was found to be far more variable than their Illumina data and was a better match to the magnitudes of variance observed in our combined Ref-8/HT-12 dataset. The justification for the low variation in the Illumina data is unknown, especially since the MAQC study design deliberately split the sample replicates before cRNA synthesis- a much earlier stage in the sample-prep workflow than our replicates (which were split after amplification and labelling). It is possible that the small number of laboratories (three, in total) performing the MAQC Illumina hybridisations produced highly concordant data completely by chance, while the larger number of laboratories (six, in total) performing the Affymetrix experiments provided a more realistic reflection of the technical variation in these data.

Analyses of the Paterson Institute and MAQC Affymetrix datasets revealed larger overall variation in the former, probably as a result of the re-extraction of RNA from culture providing an extra level of variance not interrogated in the MAQC experiments however it is likely this was slightly overestimated due to the unbalanced design of the

Paterson experiment. Overall the variance breakdown in these data are as expected, with the dominating contributions arising from different RNA extractions (samples) followed by that introduced during amplification and labelling.

4.4.1 Pool vs. UHRR

Several studies have found the use of replicate control samples such as UHRR to be a useful standard in microarray experiments, suitable for monitoring expression consistency within and across a variety of genome-wide expression platforms [202, 203, 204, 205]. However, such commercial controls are deliberately generic in terms of the RNA they contain and deficiencies have been reported in terms of their representation of more 'pure' RNA such as that derived from colon epithelial cells [206]. Similarly, a UHRR sample is not representative of breast tumour RNA and therefore carries no guarantee of expressing RNAs that may be variably expressed in the specific subset of genes changed in breast tumour tissue. Therefore, in terms of compensating for confounding technical variation, the very probes for which the correction is most important are those that are most neglected in the UHRR controls. To that end, we sought to compare UHRR to bespoke control samples derived from pools of 'representative' tumour RNA in terms of the ability of each to model the technical noise in the expression data.

As a general observation, the pooled RNA from the tumour samples used in experiment 2 picked up the same strong chip-effect found by the UHRR controls and, indeed, pairwise correlations between pool replicates were very similar to the UHRR replicates in these HT-12 data. In experiment 2 the inter-chip variation is far larger than the inter-array and inter-run variation and it is clear from the pairwise correlations between these control samples that chip 22 was the primary cause of this discrepancy. However there was nothing to suggest the samples on this chip were abnormal given the internal controls on the Illumina chips and associated quality-control analyses.

As before, array-wide quantile normalisation did not improve the agreement between the replicate UHRR or pool controls in experiment 2, but the within-batch variance moderation performed by *ComBat* successfully removed most of the effect of the outlying chip even when correcting specifically for run-level effects. This is a very encouraging result, as chip-level batch correction is an extremely severe manipulation to be performing on experiment data at such a low level and, in our opinion, run-level corrections are favourable as they are less invasive, more reliable due to greater numbers of samples per batch, and less vulnerable to preconceived notions about differential expression between sample groups.

Unfortunately the rather small number of samples in this experiment, combined

with very little legitimate biological differential expression between the pre- and post-treatment tumour biopsies in experiment 2 provided little opportunity to assess the effect of batch correction on the consistency of reported differentially expressed probes. However, compared to the UHRR, the pooled tumour RNA controls were shown to more faithfully emulate the individual shift in expression between tumour technical-duplicates as a result of variation introduced between runs and between chips. In this case, had there been more legitimately differentially expressed probes between pre- and post-treatment samples, the pooled RNA would almost certainly have made for a better batch-calibrator during *ComBat* correction than the UHRR controls. If we had used similar pools instead of the UHRR during batch-correction in the first experiment it seems reasonable to speculate that the consistency between the gene-lists, reported as significantly differentially expressed, would have been noticeably higher. A previous study demonstrated that the composition of datasets should be relatively consistent for meaningful integration and robust meta-analysis [126]. It seems from the current study that different datasets should also be of a reasonable size, although comparing identifying the minimum experiment size for combining datasets requires further work.

We also took the opportunity to assess compositional properties of the probes as a potential explanation/surrogate for the technical effects observed in our Ref-8 and HT-12 data. A highly significant trend in favour of low-GC content in the core set of probes consistently affected by inter-run and inter-chip variation between sample duplicates in experiment 2. We assumed that GC-content is likely to be a subtle effect, resolvable only in highly similar sample replicates such as ours, in which all processing steps prior to hybridisation were shared. However we were somewhat surprised to find a similar, significant, enrichment for probes with low-GC and high-SD in the MAQC Illumina dataset. This suggests that the magnitude of error introduced due to low probe GC-content is sufficiently great that it is resolvable between the replicate cRNA preparations assessed in the MAQC study; but less obvious than in samples from replicate hybridisations of the same amplified, labeled cRNA used to generate our Ref-8/HT-12 data.

Correlation of probe composition, specifically with respect to GC content, has been reported previously in a spike-in experiment using Illumina BeadChips [51]. Here it was found that probes with high-GC content tended to have a higher than expected signal intensity, but probes with lower than average GC content had inflated differential expression statistics. These findings agree with our observations, and one could suppose that the low-GC/high-SD trend is a result of the increased signal from high-GC probe-targets that are thus more likely to be reliably detected. Although this justification may not be entirely relevant in this case due to the use of technical replicates and relatively stringent detection filter to remove unreliably expressed probes. Also, analyses of

a subset of biological replicates on our Ref-8 chips did not show any meaningfully significant low-GC effect by chi-square analysis. Therefore the low-GC effect is more likely to be related to thermodynamic properties of hybridisation favouring high-GC probes/targets, a supposition that is rational given the deliberate high-GC bias in the design of the Illumina probesets.

The proximity, with respect to the target transcript, of probes has been reported to strongly influence the correlation of expression measurements between technologies [71]. The analyses performed here were designed to assess whether such probe-transcript mapping influenced expressions reported by the same platform, however no such correlation was observed either between biological or technical replicate samples. A more thorough analysis of the MAQC datasets would provide further insight into any relationship between probe-location and expression between a variety of different platforms and sample-preparation procedures. However, this is left as future work.

4.5 Conclusion

The primary source of systematic error in any given microarray experiment is unpredictable, but can generally be attributed to RNA extraction and, to some extent, labelling and amplification. However, pools of RNA derived from of clinical 'analysis' material and commercial control samples such as UHRR are effective in modelling the variation, especially when this variation is large. However pools of RNA, relevant to the investigation at hand, outperform UHRR in the extent to which they emulate the effects of systematic error, acting as a more effective calibration sample during batch-correction. Probes with low GC-content are inherently more vulnerable to systematic error, but although highly statistically significant, the magnitude of this variation in our data was small relative to that between biological replicates. Given the results presented here, including those derived from external data sources, it is desirable in a clinical context to not avoid analysis of individual test samples (to then try and classify), but instead to run a several at the same time to get a better handle on the experiment variation and be better equipped to compensate for it.

Moreover, sound experiment design is of critical importance to avoid confounding systematic variation. Randomisation of samples over the different arrays on each BeadChip, blocking of samples from each sample-group not only within each run, but within each BeadChip are the only sure-fire methods of protecting against unwanted technical noise in Illumina array data. In situations where this is not possible, detailed meta-data should be preserved for each array that includes, at the very least, the date and time of each hybridisation and scan. In either case, diagnostic procedures such as PCA or SVA should be routinely performed prior to downstream array analyses.

Chapter 5

Comparison of expression measurements and results obtained from matched RNA assessed by Illumina BeadChips and Solexa sequencing

Preface

Following on from the attempts in previous chapters to provide insight into the effect of experiment design and sample preparation on reported gene expression levels; in this chapter I further explore the variability between two very different technologies; microarrays and second generation RNA-sequencing (RNA-seq).

The majority of the results and data in this chapter are unpublished, however the findings contribute to two publications currently in press [207, 208]. Again, as was the case in previous chapters, I was not responsible for the design of the study nor any of the biological processing of the samples, including animal handling, sample collection, RNA preparation, sample-pooling, array hybridisation, or RNA-seq library preparation; these tasks were performed by Kelly A. Bordner, Arthur A. Simen, and Shrikant M. Mane. However all of the analyses and evaluation presented herein were performed entirely by myself with some direction provided by Arthur A. Simen, Mark B. Gerstein, and Andrew H. Sims.

Abstract

Background: Whilst a number of investigations have been performed to examine the reliability of gene expression measurements obtained from matched RNA assessed by different microarray technologies, relatively few studies have compared expressions derived from the more recent second-generation shotgun RNA-sequencing (RNA-seq) experiments with those obtained by microarrays. We aimed to provide a comprehensive comparison of array and RNA-seq technologies in terms of both the raw expression measurements and the results of differential expression analyses at different resolutions of the genome.

Results: Comparison of different read-mapping and quantification methods produced surprising results, with 25% of genes exhibiting variation above the level expected by Poisson noise, probably resulting from poor quality sequence reads. Correlations of both raw expressions and differential expression estimates between the technologies were relatively high and the increased sensitivity of RNA-seq, despite the poor quality of the reads, was very clear compared to the arrays. Agreement in the direction of change in reported fold-changes was excellent, but absolute fold changes were not well correlated. Consistency in raw expressions between two sets of technical replicates sequenced over two years apart was surprisingly good, with only 7% of genes exhibiting extra-Poisson variation despite poor overall quality of both sets of reads.

Conclusion: Despite concerns about data quality, we found that the choice of method for mapping and quantifying the raw sequence reads can have a large effect on the expression estimates. Overall we found generally good agreement between RNA-seq and arrays; integration of fold-changes reported by each technology is reasonable, after compensating for the overestimation of changes by RNA-seq, and the pooling of samples prior to sequencing was likely the reason for the low power, compared to the arrays, to detect differential expression. Gene- and exon-level analyses were approximately equivalent in terms of comparison with arrays and also between technical & biological replicates on RNA-seq. Analysis at the transcript-level is clearly the most desirable method, but requires advanced and careful mapping and quantification in order to provide reliable expression estimates.

5.1 Introduction

Since the mid-1990s, microarrays have been the technology of choice for large-scale, high-throughput mRNA gene expression studies. However, as discussed at length in the preceding chapters, array-based technologies have several technical limitations that limit the precision and accuracy of the expression measurements derived from them. For example variable background noise, due to random and non-specific binding of cRNA in the sample to probes on the array, is particularly disruptive for transcripts present of low abundance and defines the lower limit of detection [209, 210]; probes for different genes tend to differ considerably in their thermodynamic hybridisation properties; and a limit on the upper ranges of expression due to probe saturation [87].

Finally, and perhaps most importantly, arrays are limited to the interrogation of only those known (or speculated) genes, transcripts, and exons for which relevant probes have been deliberately designed. This design process, and indeed the very interpretation of array results, makes several assumptions about the structure and transcription of the genome that is increasingly being found to be more complex and dynamic than was previously thought [211]. Perhaps with the exception of tiling arrays [212], microarrays are not capable of identifying, for example, novel transcript isoforms, fusion transcripts, and pseudogenes [213]. In fact, detection of such events within the transcriptome of the sample under analysis would serve only to obfuscate expression measurements and complicate interpretation of the true underlying biological processes.

Sequencing-based approaches to measuring gene expression levels have the potential to overcome all of these limitations and new, ultra-high-throughput, 'second-generation', sequencing techniques enable thousands of megabases of DNA to be sequenced in a matter of hours and days [10]. Second-generation RNA-sequencing (RNA-seq) provide, for the first time, the capability to directly sequence almost the entire cDNA transcriptome(s) contained within a given sample, in a high-throughput and affordable manner. It is this ability to estimate expression across the entire length of a target transcript, rather than just within small region targeted by a probe, that makes RNA-seq such an appealing prospect compared to existing technologies such as arrays and qPCR [13].

There are a number of differences in the way in which the various second-generation sequencing platforms produce short reads that fall outside the scope of this chapter, but a comprehensive review is provided in [11]. However, as with microarrays and qPCR, all of the various sample preparation procedures [99] required for RNA-seq are equally liable to introduce noise into the experiment, serving to obscure legitimate and interesting biological variation under investigation.

The fundamental measurement obtained by RNA-seq and microarray are also very

different. Expression estimates for array probes are produced directly from their fluorescent intensity following laser-excitation, of the dye attached to bound cDNA (or cRNA) target sequences in the sample. Unlike arrays, the signal obtained from RNA-seq is essentially the reads themselves and is therefore discrete, not continuous in nature. Expressions derived from RNA-seq experiments can only be estimated following assignment of all (or as many as possible) of the millions of short reads to the genome or transcriptome of the organism [214]. This difference between continuous fluorescence and discrete read-counts leads to very different statistical considerations that impact the filtering, normalisation, and comparison of samples. For example, an array probe can be filtered based on its fluorescent intensity vs. the local background and is filtered after quantification, as opposed to an RNA-seq read which comes complete with per-base quality scores and so is filtered during mapping to the reference sequence.

Despite the promised improvements in expression detection, precision, and accuracy provided by the short-read sequencing approach, it is still a nascent technique. Many recent articles have shown progress in attempts to understand the best methods for assembling reads [104] (especially for identification of novel exon-boundaries [215, 108]), quantifying expression [107] (including isoform identification), improving expression normalisation [110] (with proper compensation for transcript length bias [112] and highly-expressed transcripts), and analyses of differential expression [113, 117, 116]. It is little wonder then, why it is the downstream analysis that is widely seen as the rate-limiting factor to more widespread use of the technology [214].

All of the above factors influence potential reliability of RNA-seq and agreement with previous methods such as arrays. To address this, several studies have already compared expression data obtained by array with RNA-seq. Bradford *et al.* focused on an exon-level evaluation of RNA-seq, in a comparison with Affymetrix exon arrays [216], finding very good agreement in the reported direction of change between cell lines, a finding that has also been corroborated on similar arrays [217]. RNA-seq has been compared with tiling arrays [218], in which 'reasonable' correlations were observed between raw signals but that RNA-Seq outperformed the arrays in exon boundary detection and dynamic range of expression. Finally, in an excellent article published in 2008, Marioni *et al.* [109] used matched RNA on Solexa sequencing and Affymetrix arrays, performing a quick comparison of absolute expressions between then technologies and a more in-depth analysis of differential expression.

5.1.1 Motivation and analysis plan

A small number of previous investigations have assessed the reliability of expression measurements obtained from matched RNA assessed by different microarray tech-

nologies, both within the context of a single experiment [186, 189] and between experiments, laboratories [163], and array manufacturers [182, 118]. The purpose of this study was to compare expressions and results obtained from Illumina BeadChips and Illumina/Solexa RNA-seq using matched RNA from mouse prefrontal-cortex (PFC). Mouse PFC is a tissue that presents a particular challenge for accurate expression measurements, on any platform, as it yields small amounts of RNA and differential expression is typically small. It is therefore desirable to embrace a potentially more sensitive technology such as RNA-seq, but it is also interesting to investigate similarities and differences in this new method to the array-based technologies it is expected to eventually replace. If nothing more than to assess how well probes on arrays such as the Illumina BeadChips represent expression at the resolution of a whole-exon or, as they are more commonly reported, a whole-gene.

Here, the effect of using different methods of mapping and quantification of Illumina/Solexa RNA-seq reads is compared in terms of reported expressions. RNA-seq is also compared with Illumina BeadChips in terms of both the raw expression measurements and the results of differential expression analyses at different resolutions of the genome. These resolutions included gene-level comparisons of RNA-seq read counts measured over the entire length of a gene compared to array expressions (averaged in cases where multiple probes target the same gene). Similarly for exon-level comparison of RNA-seq reads summed over the length of all exons targeted by the array probes. Finally at the probe-level, to assess whether RNA-seq is capable of producing reliable expression estimates within the very small (∼100nt) window targeted by the array probes.

5.2 Methods

5.2.1 Samples

Briefly, total RNA samples were obtained from the prefrontal cortex (PFC) of healthy, wild-type, mice at three ages: Adult (‘AD’; 6 months), middle-aged (‘MA’; 12 months), and old-aged (‘OA’; 18 months). Two cohorts of animals were used to generate these data, the first cohort (‘cohort 1’) contained animals obtained, pre-aged, from the US National Institute of Aging (NIA) while the second cohort (‘cohort 2’) was comprised of animals obtained from the Jackson Laboratories and aged in-house. All animals were maintained and treated in accordance with procedures approved by the Yale University Institutional Animal Care and Use Committee (protocol 2008-10975).

All RNA-seq and microarray experiments were performed at the Yale Centre for Genome Analysis (YCGA) [219]. RNA was extracted using the Qiagen AllPrep mini

kit, according to manufacturer's instructions. Sample quality was verified using Agilent 2100 Bioanalyzer RNA chips and in all cases the RNA integrity number (RIN) exceeded 8.0.

For the microarray analysis, both cohorts were processed at the same time and hybridised to Illumina MouseWG-6 BeadChips. Financial limitations initially restricted RNA-seq analysis to being performed only using animals from cohort 1. RNA samples from each animal in cohort 1 were split prior to cDNA synthesis, a protocol that differs between the Illumina microarray and RNA-seq experiments. Unused RNA was stored in RNAlater at -80^{o}C. While all samples were run on individual Illumina arrays, due to the high cost of sequencing, samples were pooled prior to RNA-seq analyses. Each pool contained, equal amounts of RNA from each of five animals, of equal age, from the same cohort. Composition of these pools is shown in Figure 5.1, where AD_{1-5} were combined to create ('==') AD_{p1}, MA_{6-10} == MA_{p2}, OA_{11-15} == OA_{p3}, etc.

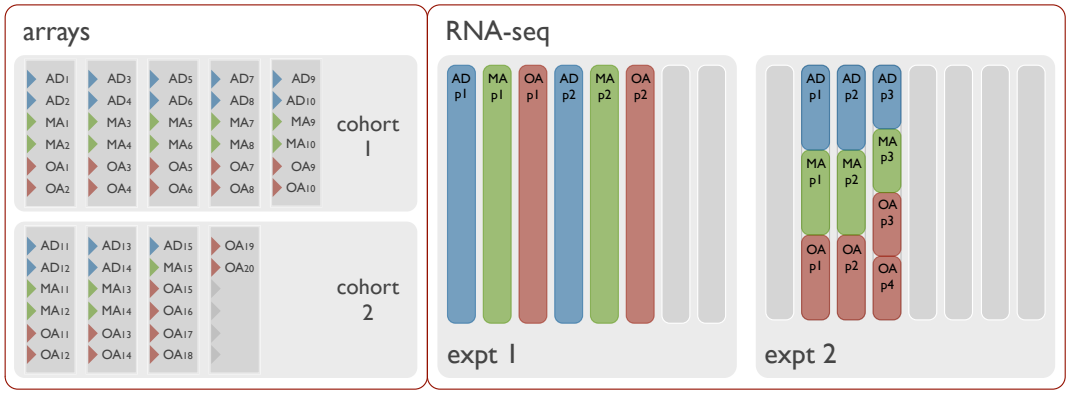


Figure 5.1: Illustration of the samples from both cohorts of animals as they were processed on Illumina BeadChips and by Illumina/Solexa RNA-seq. AD represents adult animals, MA are middle-age, and OA are old-age. On RNA-seq, pools are denoted by the trailing 'p1' to 'p4', where the third and fourth pools were created from animals in the second cohort.

Due to reasons described in the Results section, the pools from cohort 1 were recently re-analysed by RNA-seq in a second experiment ('experiment 2'; Figure 5.1). RNA from each of the individual animals was thawed, re-assessed for quality and re-pooled prior to library creation. All individual samples passed the same RIN cutoff as before. Two of the eight available lanes on the Illumina Flowcell were used for these samples, with each lane containing a pool of each age group; this was achieved by 'barcoding' each pool to allow multiplexing. Pools were created using exactly the same constituent samples as in first experiment. Also assessed in this second RNA-seq experiment were pooled samples from animals in cohort 2 using a single lane on the

Flowcell.

Again, the overall design of this study is illustrated in Figure 5.1, however the reader is referred to Bordner *et al.* [207] for complete details of animal handling and array/RNA-seq sample prep.

5.2.2 Statistical Methods

Array

All data were generated using Illumina MouseWG-6 (v.1.1) BeadChips. Probe expression and detection data were output from Illumina's *BeadStudio* software; all subsequent analyses was performed using *Bioconductor* [175] algorithms implemented in the statistical programming language, *R* (v.2.12.1) [147].

All BeadChip data were filtered, where specified, using the detection confidence reported by *BeadStudio*, determined for each bead based on the expressions of internal control probes, local background intensity, and the uniformity of the reported intensity of the bead. The filtering was performed prior to quantile normalisation such that probes with a detection confidence less than or equal to 95% in more than 20% of the samples were removed from further analysis. *ComBat* [127] batch correction was used to compensate for the effect of cohort where specified.

Analyses for differential expression between age-groups was performed using *limma* (v.3.6.9) [66] and *SAM* (v.1.28) [67]. For the analysis using *limma*, genes were defined as being differentially expressed after satisfying a minimum fold-change of ± 1.5 and a maximum, Benjamini-Hochberg adjusted p-value of 0.01. For the *SAM* analysis, the differentially expressed genes were selected at a maximum predicted false discovery rate (FDR) of 5% and the same minimum fold-change of ± 1.5.

To facilitate the comparison with data obtained by RNA-seq, all probes were mapped to the NCBI mouse transcriptomic reference (v37) [198] using *Maq* (v.0.7.1) [106], setting all nucleotide qualities to the maximum value, allowing for 0 mismatches in the alignment, and reporting only probes with unique hits to the transcriptome.

RNA-seq

Casava (v1.0) expression counts were obtained using the output from Illumina's *GERALD* analysis module (Illumina's Pipeline v1.5). Briefly, this module performs the read alignment using Illumina's 'Efficient Large-Scale Alignment of Nucleotide Databases' (*ELAND*) algorithm to the NCBI mouse genomic reference (v37). Read counting is performed by *Casava* by sorting reads with respect to their reported alignment to the genome, summarising each read as a single integer value corresponding

to the reported position of the first nucleotide, and summing the number of these first-bases within each given reference region such as a gene or an exon.

In addition to *Casava*, raw reads were also mapped to the NCBI mouse transcriptomic reference (v37) using *Maq*. Parameters were left at the default, allowing for a maximum 2-base mismatch in the alignment of the seed. Reads with multiple alignments were decided by overall mapping quality (a cumulative function of base-quality and reference match for each base in the read) and those with the same overall mapping-quality were assigned to one of the candidate reference locations at random. Custom Java software was written to summarise these *Maq* alignments in terms of the number of reads mapping to known genes, exons, and Illumina microarray probe-regions. For the probe-region quantification, expressions were estimated using only those reads overlapping the 50nt region by at least 20 bases.

Quantification of expression following *Maq* alignment was similar to *Casava*, reporting the number of reads mapping to each region, although we defined read-positions by their mid-point instead of their first nucleotide. It is assumed that this reduces bias in counting exon-spanning reads, as well as providing a little leeway at the extreme ends of the gene so as not to discard reads starting slightly before the defined gene-start site in the reference. In addition to read-counts, normalised values were calculated for each gene and exon that compensate for different sequence-yields between samples and potential biases due to differences in the lengths of genes or exons; this normalised measure of expression is the so-called 'Reads Per Kilobase of exon model, per Million mapped reads' (RPKM) [107].

All downstream analyses of BeadChip data, including mapping and expression quantification data were performed using *Bioconductor* packages in *R*. Prior to differential expression analyses, RNA-seq count data were filtered using an iterative approach with Fisher's exact-tests to find the minimum number of reads, given two libraries of different size, which are required to reach a significance of $p < 0.01$ if all of these reads were recorded from a single library and none from the other. Analyses for differential expression between age-groups was performed using *edgeR* (v.2.0.3) [117] and *DESeq* (v.1.2.1) [116]. Both methods model count data as negative binomial distributions and employ an empirical Bayes procedure to moderate the degree of over-dispersion across the features. The *edgeR* method was used with tag-wise dispersion and the smoothing parameter, prior.n, set to 2. For the *DEseq* analysis, effect size and variance parameters were estimated from the data using default parameters. In both cases, library sizes were defined as the total number of mapped reads in each sample.

5.3 Results

5.3.1 RNA-seq QC and mapping

Quality control of the six lanes of RNA-seq data immediately revealed an unusual phenomenon at bases 34 through 36 where, in the vast majority (∼80%) of the (97 million) 50bp reads, the base was called 'N'. An example of this effect is plotted, for one of the samples, using the *FastQC toolkit* (v.0.9.0; Figure 5.2). This effect was common between all samples and it was not possible to obtain a certain diagnosis based on the data, however failure of the hardware during these cycles is the most probable reason (personal communication, Mark Blaxter). If, indeed, technical failure was the cause of the poor quality bases then various subsequent issues could affect subsequent base calls and so the decision was made to trim the ends of all reads such that the total length of all reads for all samples was 32 nucleotides. This decision was also motivated because of the use of the *Maq* toolkit, in which base-qualities are used in the assessment of mapping-quality and so inaccurate base-qualities can lead to inaccurate mapping quality, regardless of position within the read [106].

The trimmed reads were then mapped to the transcriptome, as described in methods, and this *Maq*-mapping consistently assigned around 20% more reads than the untrimmed reads mapped using *Casava* (summary provided in Table 5.1). Array probes were mapped in exactly the same way and an annotation file containing their mapping position, gene, exon, and position, was created. Mapped RNA-seq reads were counted, and RPKM values calculated, using bespoke Java software (available on request) and summarised in terms of both the standard NCBI v37 reference and the reduced annotation provided by the mapped array probes. Therefore, for the RNA-seq data, expression values were provided for all RefSeq genes and exons, as well as the subset of genes and exons interrogated by the microarray probes.

5.3.2 Expression comparison: different mapping methods and technical replicates

Unlike *Maq*, *Casava* was run following the standard *GERALD* pipeline provided by Illumina using untrimmed reads (see methods). To investigate the effect of the choice of mapping method on the reported expressions we selected gene-level expressions estimated from reads obtained from lane 1 ('AD_{p1}'; Figure 5.1). The gene-level expression measures were chosen to avoid issues arising due to the presence of exon-spanning reads that would complicate the comparison.

Although both *Casava* and *Maq* read-mapping pipelines were performed using the same reference build, the former used the genomic sequence while the latter used the

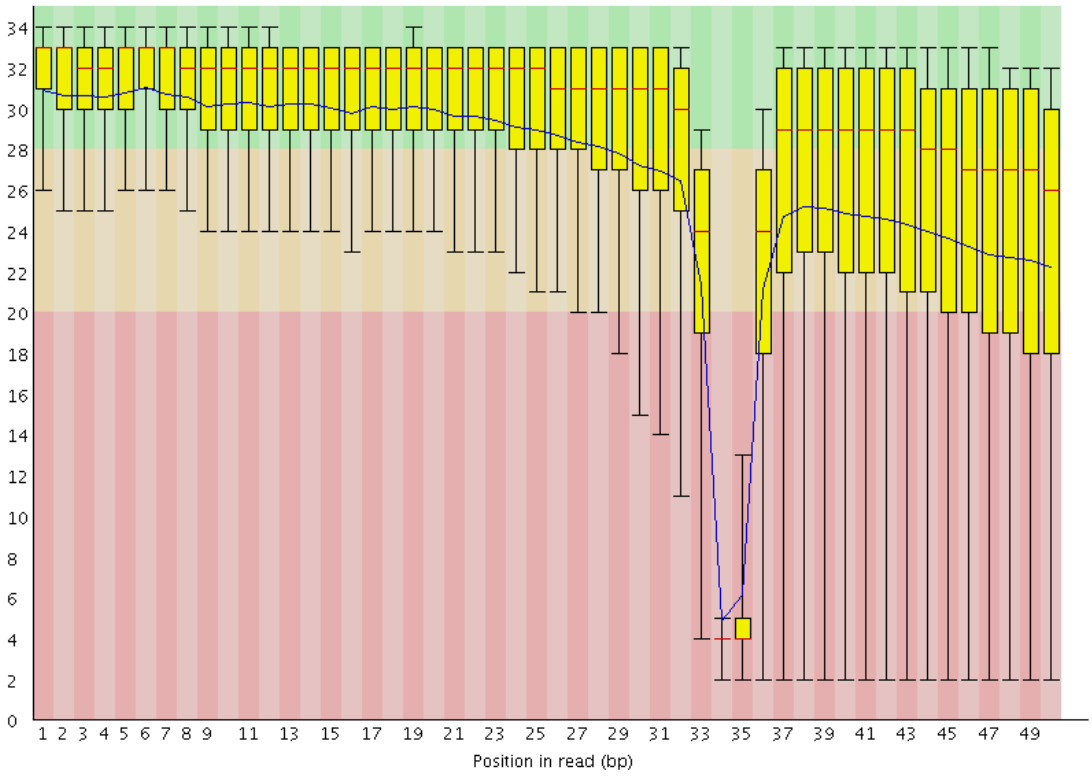


Figure 5.2: Distributions of base qualities reported by the sequencer (y-axis) for each position along the read (x-axis), over all reads in sample 'AD_{p1}'. Whisker extremes denote the 10% and 90% points, blue line represents the mean. Green, orange, and red regions denote high, intermediate, and low base quality, respectively

Lane	Total number of reads	Mapped reads - *Maq* # (%)	Mapped reads - *Casava* # (%)
AD_{p1}	14,870,439	9,770,087 (65.7)	7,133,764 (48.0)
MA_{p1}	16,191,285	11,092,235 (68.5)	7,787,219 (48.1)
OA_{p1}	16,027,933	10,842,160 (67.6)	7,550,922 (47.1)
AD_{p2}	16,360,026	11,862,737 (72.5)	8,209,949 (50.2)
MA_{p2}	17,244,988	12,798,551 (74.2)	8,426,676 (48.9)
OA_{p2}	16,517,620	11,843,920 (71.7)	8,212,139 (49.7)

	Total number of probes	Mapped probes - *Maq* # (%)
Array	45,280	35,025 (77.4)

Table 5.1: RNA-seq read- and array probe-mapping summary

transcriptomic sequence (see methods). The two methods also used different gene IDs and so the UCSC *knownGene* database [220] was used to harmonise the two sets of expression measurements, translating accession numbers to gene symbols. For simplicity, genes with multiple accession-numbers or gene symbols were excluded from the comparison (leaving 15,610 genes for analysis).

A comparison of read-counts reported by each mapping method is provided in Figure 5.3. All counts were increased by a fixed value of 0.25 prior to taking the logarithm so as not to exclude genes with no mapped reads, however the 3,171 genes with no mapped reads in both *Maq* and *Casava* were removed from the plot to better resolve the density gradient in the remaining points. There is a clear trend about $x \sim y$ (Pearson correlation = 97.3%) between the different quantifications methods, although a large number of genes are reported as expressed in one method, while not in the other.

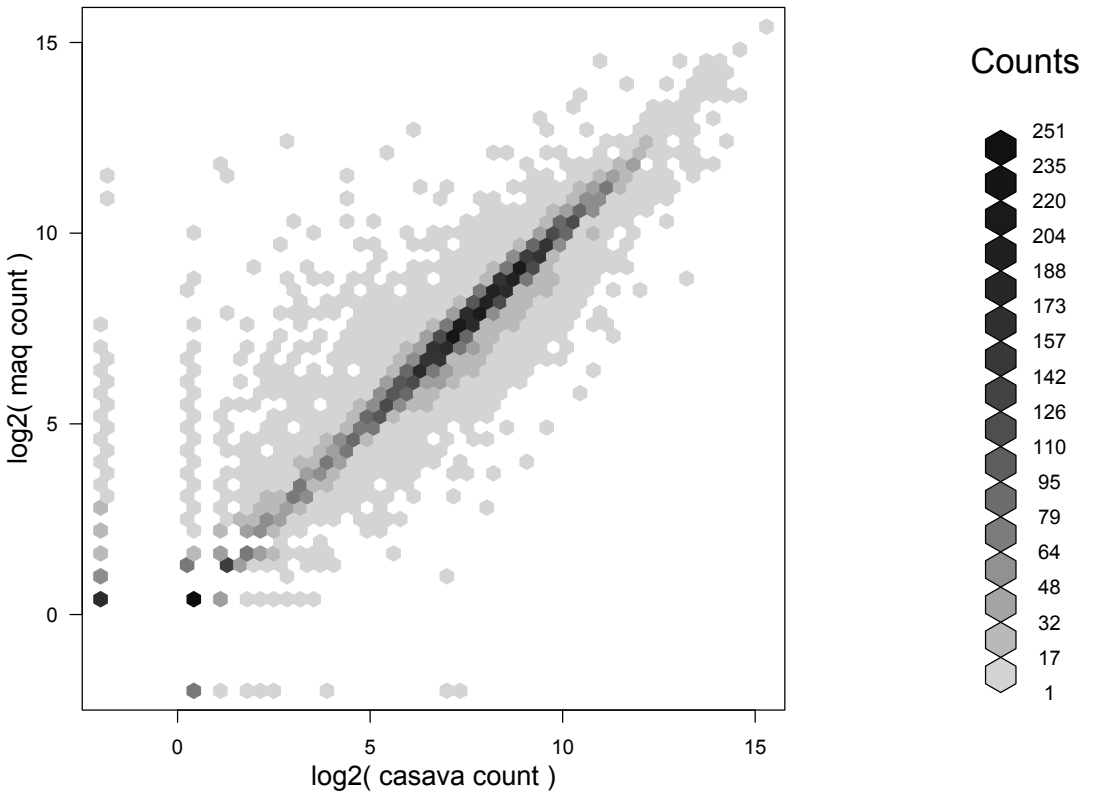


Figure 5.3: Gene-level comparison of raw expressions estimated by Illumina's *ELAND/Casava* (x-axis) and our custom-counted *Maq*-aligned reads, trimmed to include only the first 32 bases (y-axis). Genes with zero counts reported by both *Casava* and *Maq* are excluded from the plot so as to better resolve the remaining points.

To put this in context we analysed a second RNA-seq dataset, described in [208] but not used in any further analyses in this chapter, in which three pairs of replicate library-preparations were run on a single Solexa chip. Using *Maq*, we performed exactly the same mapping as described in Methods and plotted normalised expressions (Reads Per Kilobase of exon-model per Million mapped reads; RPKM) for each of the replicate pairs at the gene-, exon-, and probe-level (Figure 5.4A - 5.4C, respectively). Pearson correlations between the replicates at the gene-level was 99.6% and almost identical at the exon-level (99.5%). Slightly less good at the probe-level (96.4%), due to the much smaller reference region in which the reads were to be mapped, but nevertheless a similar correlation to that between replicate arrays seen in previous chapters. Similar distributions and correlations were obtained from analyses of count data (data not shown). To put these correlations in context, we also analysed expressions between 18 pairs of technical replicates used in a recent publication by Blekhman *et al.* [111]. In this dataset we found that the average gene-level Pearson correlations were 99.987%, which is 30-times greater than the gene-level correlations observed in our data.

For each pair of technical duplicate samples we computed, for each gene or exon, a p-value based on the null hypothesis that the difference between reported read counts could be explained by random re-sampling of the reads in both samples. Significances were obtained by two-tailed Fisher's exact tests, in a similar manner to [109], and we found only a small fraction of the total genes (1.03%) and exons (0.35%) exhibited clear evidence ($p < 0.01$) of non-random deviation between the duplicate samples. At the probe-level the variation was much higher, with 18.41% of the features showing significant departure from the Poisson noise model, suggesting that analyses at this level are too variable to be of use in assessing differential expression. Plots of observed vs. uniform quantiles for the p-values revealed deviation from uniformity only for highly significant ($p < 0.001$) genes and exons (data not shown).

Similar analyses using Fisher's exact tests found that just over a quarter of the genes in the *Maq* vs *Casava* read-counts comparison exhibit non-random variation, much larger than the Poisson-noise observed between pairs of technical sample replicates. This therefore suggests that accurately mapping reads to the genome / transcriptome is potentially of much greater importance than running such replicates.

There are several possible reasons for this variability (shown in Figure 5.3) including the fact that alignment by *Casava* and *Maq* employ different methods of counting reads (read-start vs. read-centre, respectively) and possibly handle overlapping genes differently. Perhaps the most compelling potential justification however, is that from exactly the same input data, *GERALD/Casava* only mapped 7,133,764 reads, compared to 9,770,087 reported by *Maq* (Table 5.1). Perhaps the trimmed reads used in the *Maq* mapping allowed more to be aligned as there is less chance of failing to map

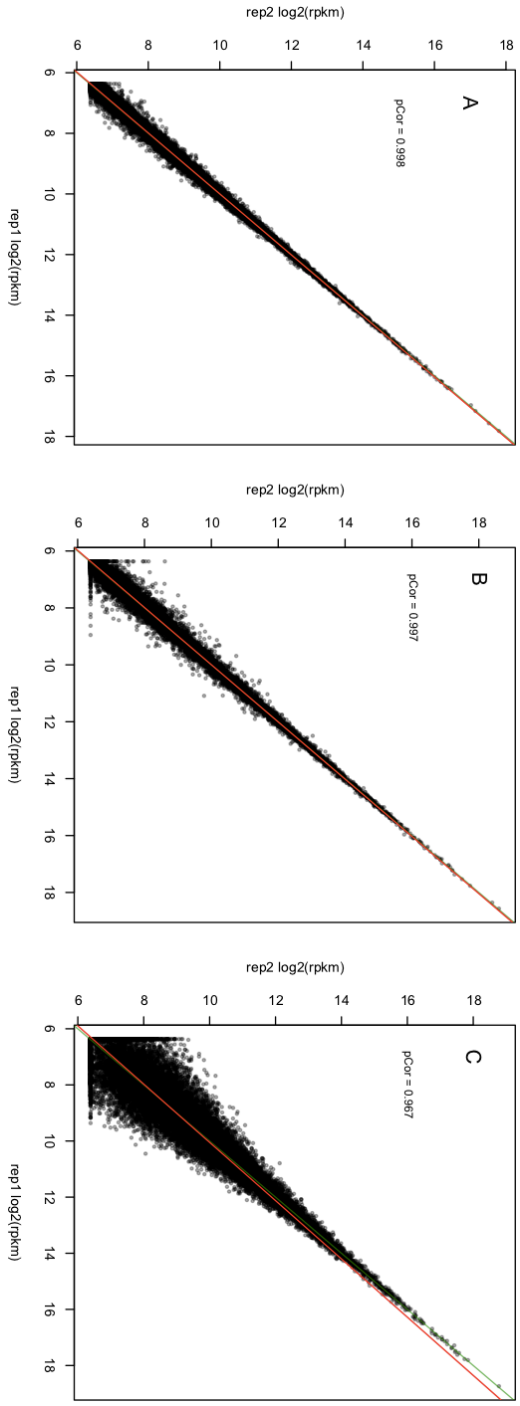


Figure 5.4: Comparison of a pair of technical sample-duplicates at various levels of the genome, produced from a different RNA-seq experiment, still using mouse brain tissue, and again mapped using *Maq*. Red line is the result of a linear regression and the green line is zero-intercept, unit gradient. Pearson correlations, pCor, are shown for each comparison. **A**: gene-level expression comparison, **B**: exon-level expression comparison, and **C**: probe-level expression comparison

reads based on the poor quality / N bases observed in Figure 5.2.

5.3.3 Expression comparison: RNA-seq vs. arrays

Using the subset of RNA-seq data mapping to the same genes and exons interrogated by the arrays, a comparison of RNA-seq expressions with array expressions reveal relatively high concordance between the technologies, regardless of the choice of mapping method. Array probes mapping to the same 15,610 genes used in the comparison of *Maq* and *Casava* were plotted against the reported *Casava* count (Fig 5.5A), *Maq* counts (Fig 5.5B), and *Maq* RPKM (Fig 5.5C). In each plot more than half of the genes were below the detection threshold on the array (8,700). Correlations used only those genes detected on the arrays and those that had more than zero counts by RNA-seq.

Despite observed differences in expressions reported by different mapping methods, at the gene-level, *Maq*-quantified counts were only slightly better correlated with array expressions than the *Casava* counts (52.8% vs. 49.3%, respectively), however the greatest improvement came from the use of normalised RPKM expressions (60.8%). Very similar correlation was observed at the exon level (53.8% and 59.0% for *Maq* counts/RPKM, respectively) and the trend is the same (data not shown).

Plots similar to these have been shown in several previous papers [87, 109], however expressions derived from RNA-seq and arrays have very different distributions (Figure 5.6) especially following the removal of unreliably detected or low expressing probes/genes (green). Using normalised RPKM expressions does make the distribution slightly more similar to the array, potentially explaining the improved correlation.

To overcome the limitations of the fundamentally different expression distributions between arrays and RNA-seq, for each technology we ranked the 6,910 genes detected by array (above the red line in Figure 5.5) such that genes with low expression had low rank and those with high expression had high rank. Ordered by increasing RNA-seq gene-rank, array ranks were plotted against RNA-seq ranks reported by *Casava* count (Fig 5.7A), *Maq* counts (Fig 5.7B), and *Maq* RPKM (Fig 5.7C). Plotting against the rank distributes the data evenly along the both axes and thus facilitates the visualisation of expression heterogeneity. In all three cases it is clear that the strongest agreement between the platforms is for the lowest and the highest expressing genes. Also interesting is the systematic effect causing the bowing of the array points compared to the RNA-seq ranks; such a relationship might infer a sigmoid-like trend in the absolute expressions, however attempts to fit such a model to the data (above the array detection threshold) in Figure 5.5 were influenced too heavily by the large number of highly-expressing RNA-seq, low-expressing array genes and did not produce satisfying results.

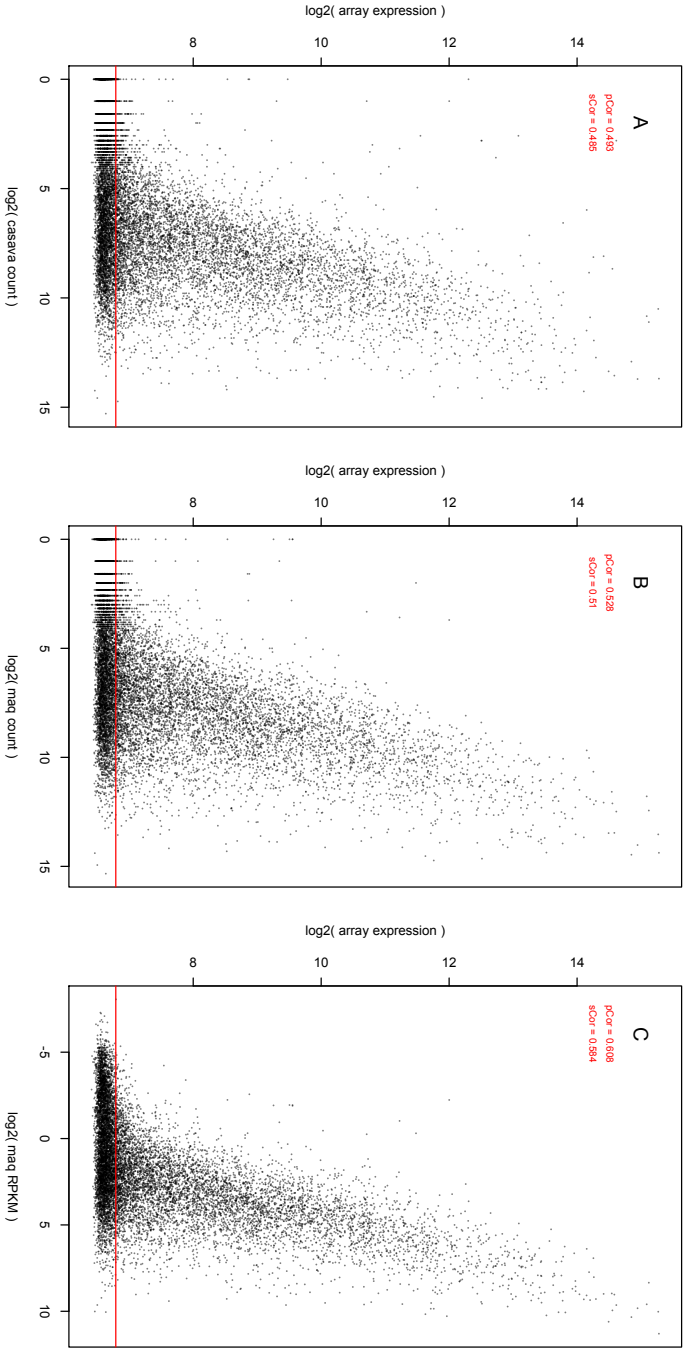


Figure 5.5: Comparison of absolute, gene-level, expressions between RNA-seq (x-axes) and array (y-axes). In cases where multiple probes mapped to the same gene, the mean of their expressions was used. The red line illustrates the approximate detection-limit of the arrays and correlations (both Pearson, 'pCor', and Spearman, 'sCor') were calculated using only the 6,910 genes above this detection limit. Against the same array-expression data are plotted RNA-seq expressions as reported by **A**: *Casava* counts, **B**: *Maq* counts, and **C**: *Maq* RPKM.

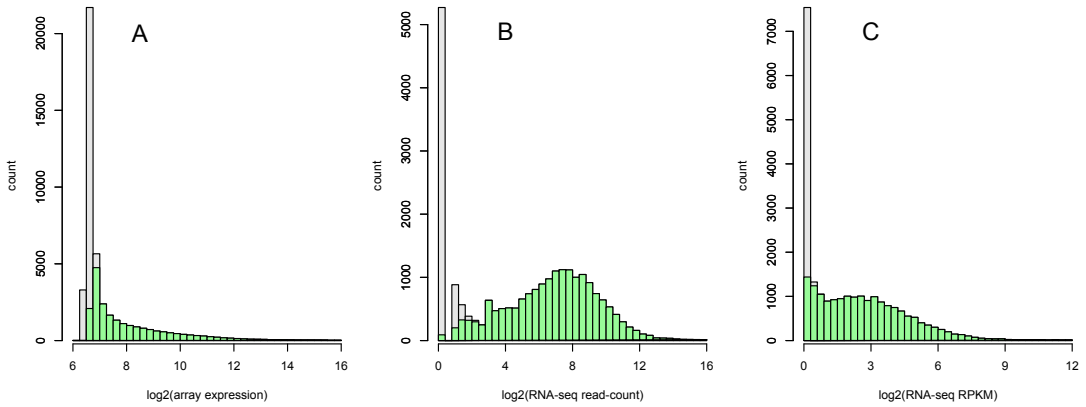


Figure 5.6: Expression distributions- grey bars indicate all probes/genes, green bars are the probes/genes that remain following detection filtering. **A**: Array expression distribution for all probes, **B**: RNA-seq (*Maq*) counts, and **C**: RNA-seq (*Maq*) RPKM.

5.3.4 Comparison of differential expression estimates by arrays and RNA-seq

Given that the main strength of arrays is in the reporting of relative expression differences between samples, rather than absolute mRNA quantification, we assessed the array and RNA-seq datasets for differential expression between pairs of each of the three age groups. The increased number of mapped, trimmed reads and the availability of normalised, RPKM expression measurements at both the gene- and exon-level led to the decision to perform the differential expression analyses using only the *Maq*-mapped data.

Principal components analysis (PCA) on the detection filtered, quantile normalised array expressions revealed a large amount of variability between biological replicates within the three age groups (Figure 5.8A), with the first two components accounting for around 20% of the total variance; despite the high variability there is a reasonable separation of samples according to age. PCA and hierarchical clustering performed on the filtered RNA-seq (gene-level RPKM) samples also revealed high variability between the biological replicate pools, with none clustering as expected (Figure 5.8B-C). The same effect was also observed using exon-level RPKM expressions (data not shown). This variability was reflected in the results of differential expression analyses on the array data using *limma* (see methods), which reported very few significant probes with a FDR corrected $q < 0.05$. Similarly with RNA-seq, analyses using *edgeR* and *DESeq* (see methods) reported no significant features that survived FDR adjustment at either the gene- or exon-level.

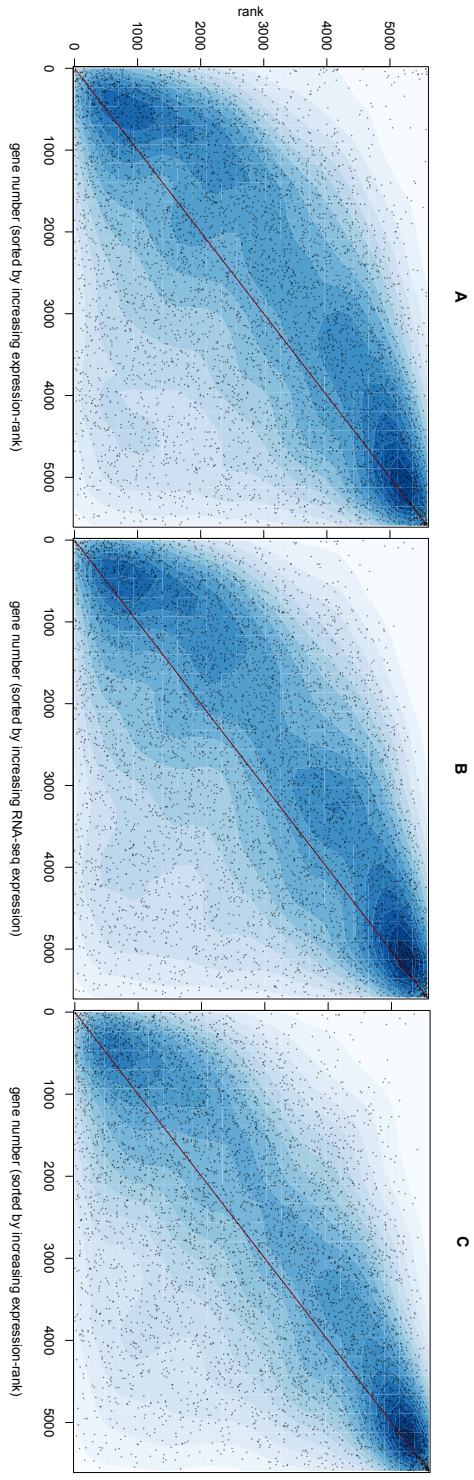


Figure 5.7: Array expression ranks (blue points) plotted against increasing RNA-seq expression rank (red line) derived from: **A**: *Casava* counts, **B**: *Maq* counts, and **C**: *Maq* RPKM. Gradient corresponding to density of array ranks is overlaid as a visual aid.

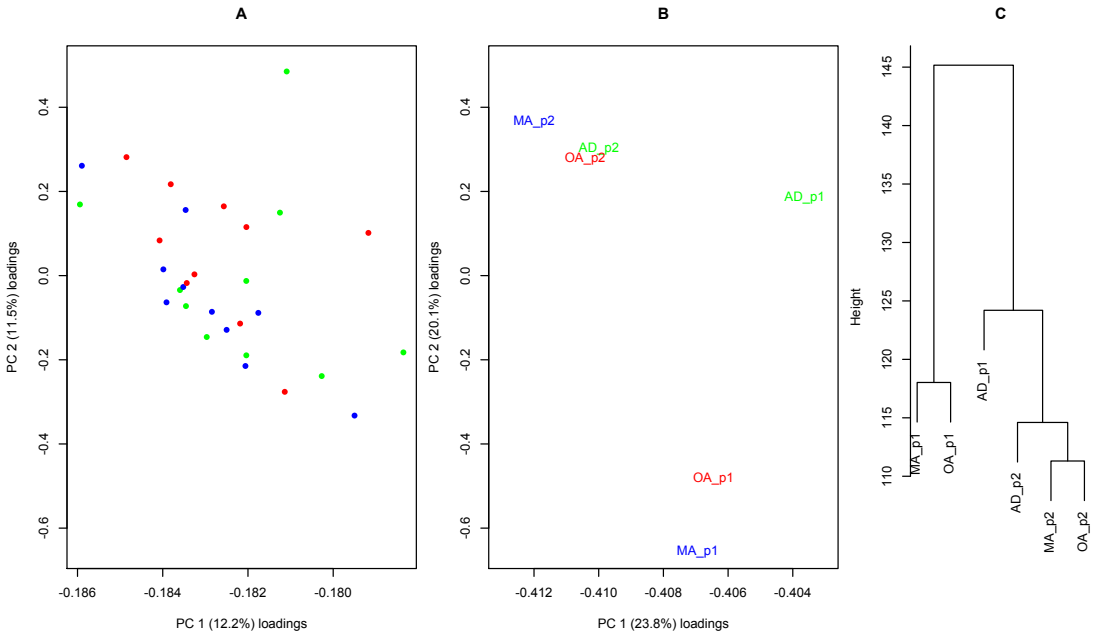


Figure 5.8: **A**: Scatter-plot of the loadings of the first two principal-components of filtered and normalised microarray samples show slight separation of samples corresponding to age. Points are coloured according to animal age: green AD, blue is MA, and red is OA. **B**: Scatter-plot of the loadings of the first two principal-components of filtered and normalised RNA-seq data. **C**: Basic cluster-dendrogram of the filtered and normalised RNA-seq data.

Despite the poor yield of reliably differentially expressed features, we compared differential expression estimates between the technologies using the reported fold-changes between age groups. In general, a positive correlation exists between RNA-seq and array (Fig 5.9A). Filtering genes such that only those with $p < 0.05$ in both the arrays and on RNA-seq remain, the result is not particularly convincing in terms of direct correlation between the magnitudes of the fold-change between the technologies, but are highly similar in terms of their predicted direction of change; i.e. if arrays report positive fold-change between age groups then, chances are, so does RNA-seq (χ^2 pVal $= 1.3 * 10^{-46}$; Fig 5.9B). Analyses at the exon level revealed similar results, with the regression of all RNA-seq and array expressions yielding an almost perfect unit gradient and zero intercept (Fig 5.9C) and, following filtering for features with $p < 0.05$, again good correlation between reported direction of change (χ^2 pVal $= 2.7 * 10^{-21}$; Fig 5.9D).

In all analyses, RNA-seq consistently reported far fewer consistent ($p < 0.05$) genes and exons compared to arrays. In order to attempt to understand this disparity, we performed gene-level RNA-seq differential expression analyses, using *edgeR*, on all possible sample permutations (Figure 5.10). From these permutations it is clear that plenty of genes are called differentially expressed whenever certain samples are combined in the same group. Specifically, in any analyses in which lanes 2 and 3 are in the same group, more than 3,000 genes are reported to have $p < 0.05$ and more than 1,000 genes have an FDR adjusted $q < 0.05$. This indicates that *edgeR* is perfectly capable of detecting differential expression in these data, and that the FDR adjustment works as expected. It is simply the case that biological variability between the 'true' age-replicate samples is too high, probably due to the pooling [221].

5.3.5 Follow-up RNA-seq experiment

Given the uncertainty in the quality of the raw reads, the poor clustering compared to array, and the consistently more conservative estimation of significance levels from differential expression the RNA-seq pools were re-sequenced; the design of this follow-up experiment ('experiment 2') is illustrated in Figure 5.1.

There was some concern that some of the variability between biological replicates run in the first RNA-seq experiment might have been caused by technical variability between sequencing lanes. Although the likelihood of a lane-specific effect contributing sufficient noise to affect the analyses of differential expression is small, in addition to other studies, our own data (Figure 5.4) demonstrated that the variation between technical replicates on the same chip is approximately Poisson [109, 111]. To be on the safe side, samples were multiplexed using the Illumina 'barcoding' protocol such

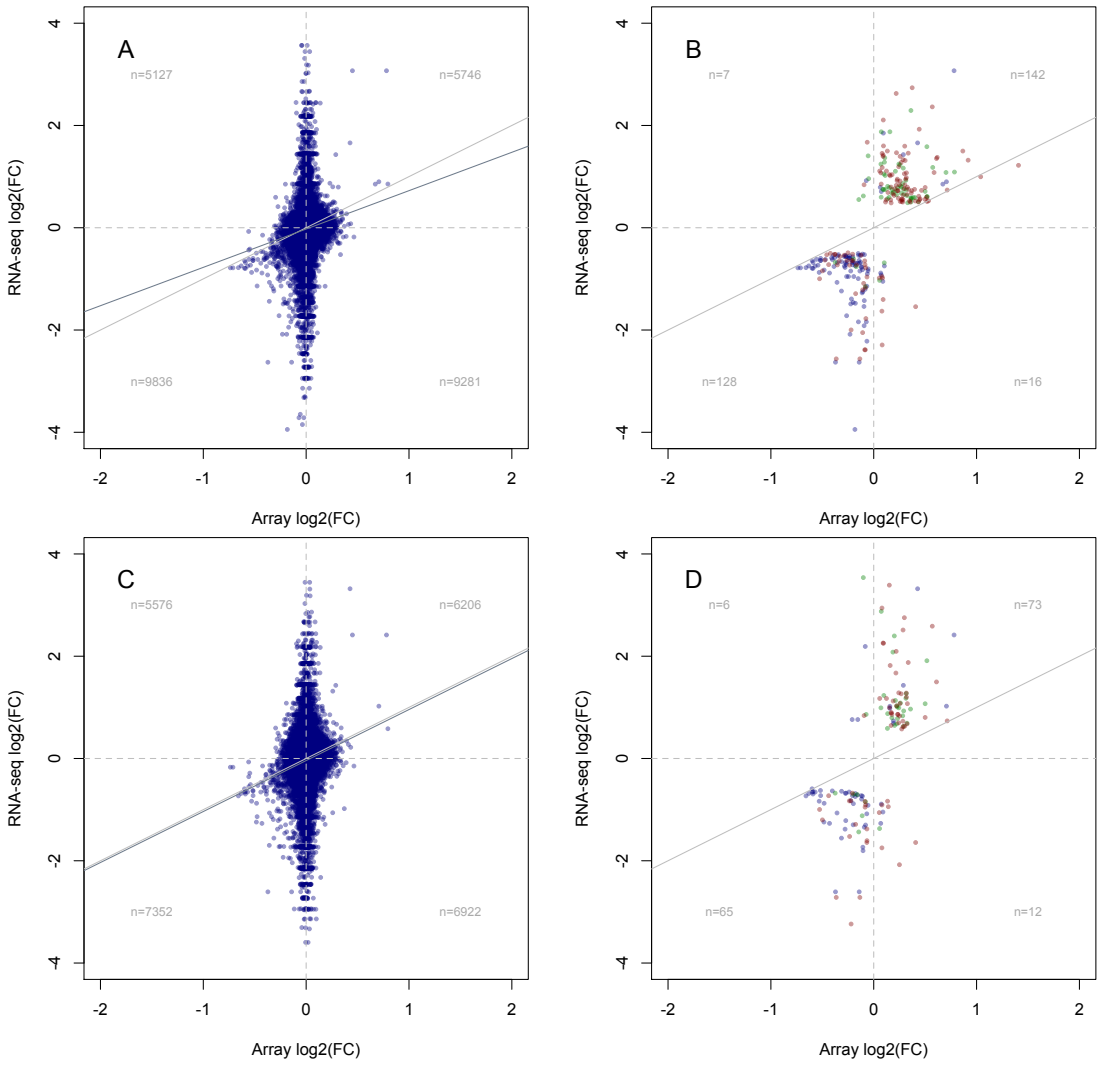


Figure 5.9: Log base-2 fold-changes as reported by RNA-seq (y-axes) and microarray (x-axes). For convenience, in each plot the number of points in each quadrant is provided, as are lines of unit gradient coloured grey. **A**: All genes in the only the single comparison of MA vs. AD age groups, the blue line is the linear regression summarising bulk distribution of points near the origin. **B**: Plot of genes in all three pairwise comparisons between the age groups; MA vs. AD in blue, OA vs. MA in green, and OA vs. AD in red. Data have been filtered such that the plot contains only genes for which the reported p-value is less than 5% in both array and RNA-seq in any of the three contrasts. **C**: Same as **A**, but using instead all exon-level differences between the MA and AD age groups. **D**: Same as **B**, again using exons with reported p-value less than 5% on both technologies in any of the three comparisons.

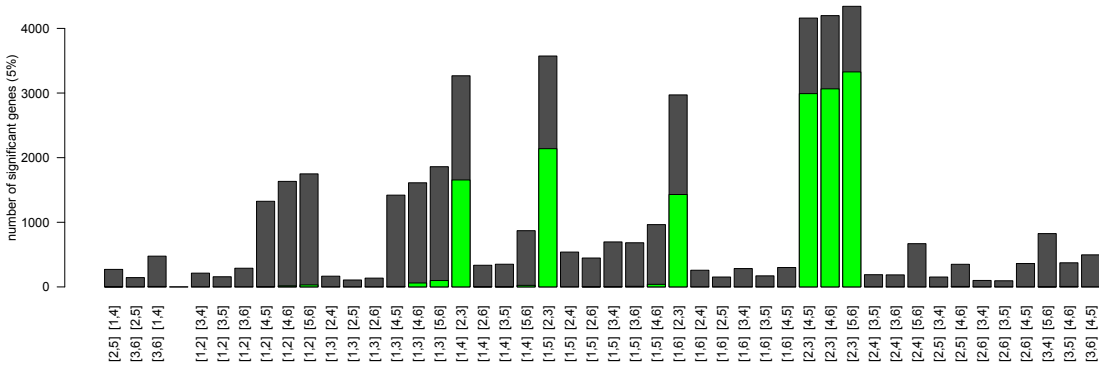


Figure 5.10: Number of genes with $p < 0.05$ reported by *edgeR* in all possible sample pairs. Sample IDs (1-6) correspond to the lane on which each sample was run (see Figure 5.1; AD=1 & 4, MA=2 & 5, OA=3 & 6). Sample pairs corresponding to the 'correct' grouping of biological replicates are shown at the left hand side. Number of significantly differentially expressed genes (for which $p < 0.05$) is illustrated by the black bars and the number still significant following FDR adjustment ($q < 0.05$) shown by the green bars.

that all six pooled-samples were run over two lanes and each lane contains one pool from each age group. In addition to the re-sequencing of the original samples, a third sequencing lane was used to sequence 4 pools of samples from a second cohort of aged animals that was interrogated using arrays at the same time as the animals from the first cohort (see Figure 5.1).

A diagnostic assessment of all samples was performed for comparison with the quality of the reads from the first round of RNA-seq. Distribution plots of base-quality over the length of all reads was again the most informative, revealing that the quality of the repeat sequenced samples from cohort 1 was very poor, perhaps as a result of long term RNA storage and numerous freeze-thaw cycles (Figure 5.11 - top). Fortunately, the per-base quality-distributions observed in reads obtained from the new cohort 2 pooled-samples was much higher (Figure 5.11 - bottom). However a much lower total read-yield was delivered in the new sequence samples ($\sim 9 * 10^5$ reads for each sample) and the number of reads successfully mapped with *Casava* ranged between $1 * 10^5$ and $8 * 10^5$, more than an order of magnitude smaller than the $\sim 8 * 10^6$ *Casava*-mapped reads per sample in the first dataset (Table 5.1). All of the second RNA-seq reads mapped using *Casava* (v1.3) which, by default, reported expression levels based on the number of nucleotides mapping to a feature, allowing proper calculation of normalised RPKM expressions.

The combined microarray data over both cohorts exhibited a noticeable batch effect, which was effectively removed by *ComBat* [127] (Figure 5.12, top row). It

Figure 5.11: Distributions of base qualities (y-axes) reported by the sequencer for each position along the read (x-axes), over all reads in **Top**: sample 'AD_{p1}', re-run in the second RNA-seq experiment and **Bottom**: sample 'AD_{p3}', a fresh pool of RNA from the second cohort of animals. The whisker extremes, blue line, and colour-scheme is the same as those used in Figure 5.2.

was difficult to see any meaningful separation between age-groups, before or after batch-correction, in either the combined cohorts or just the second cohort due to high biological variability. RNA-seq samples from the second cohort look slightly better, in that the two biological replicate OA samples cluster quite closely. However with no other biological replicates in any of the other age groups, it is impossible to make any reliable statements about the extent of biological variation in these data. Even following RPKM normalisation, RNA-seq samples from animals of the same age do not cluster together across the cohorts in either gene- or exon-level data (Figure 5.12, middle and bottom rows).

Differential expression analysis on these new RNA-seq data revealed that a small number of genes (N=38) in cohort 1 survived FDR adjustment (although no exons) and a larger number of significant genes and exons in cohort 2 (N=101 and N=281, respectively). Array analysis of cohort 2 samples reported no probes significant following FDR adjustment, but a few from the *ComBat*-corrected combined dataset.

Comparison of replicate RNA-seq *Casava* quantifications

In addition to quantification by read-bases, we also performed quantification based on read-counts for comparison with the output from *Casava* v1.0 in the first experiment. Read-count expression comparison in cohort 1 between old RNA-seq run and the new re-sequenced data perfectly follows the predicted gradient (normalised to compensate for the large differences in the total number of mapped reads) in both gene- and exon-level data (Figure 5.13). Correlation is better at the exon-level due to the large number of features with no mapped reads. Given the uncertainty over base-quality and the large difference in the total/mapped number of reads in the old compared to the new data; the correlation is surprisingly high.

As for the technical duplicate samples used in Figure 5.4, we again computed, for each gene and exon, p-values from two-tailed Fisher's exact tests to explore whether differences in counts between the old and the new RNA-seq samples could be explained by random re-sampling. Compared to the technical duplicates run on the same chip, we found a much larger fraction of the genes (7.47%) and exons (1.31%) exhibited clear evidence ($p < 0.01$) of non-random deviation between the old and the new samples. The latter, again, benefiting from a large number of exons for which no or very few reads mapped in either dataset.

Comparison of the two counting modes of *Casava*, 'read-start' and 'read-bases', using the newly sequenced samples shows the bias in reported expression introduced by the former- especially when looking at exon-level data (Figure 5.14). Little difference is observed at the gene-level, which is reasonable as few reads are expected to fall outwith

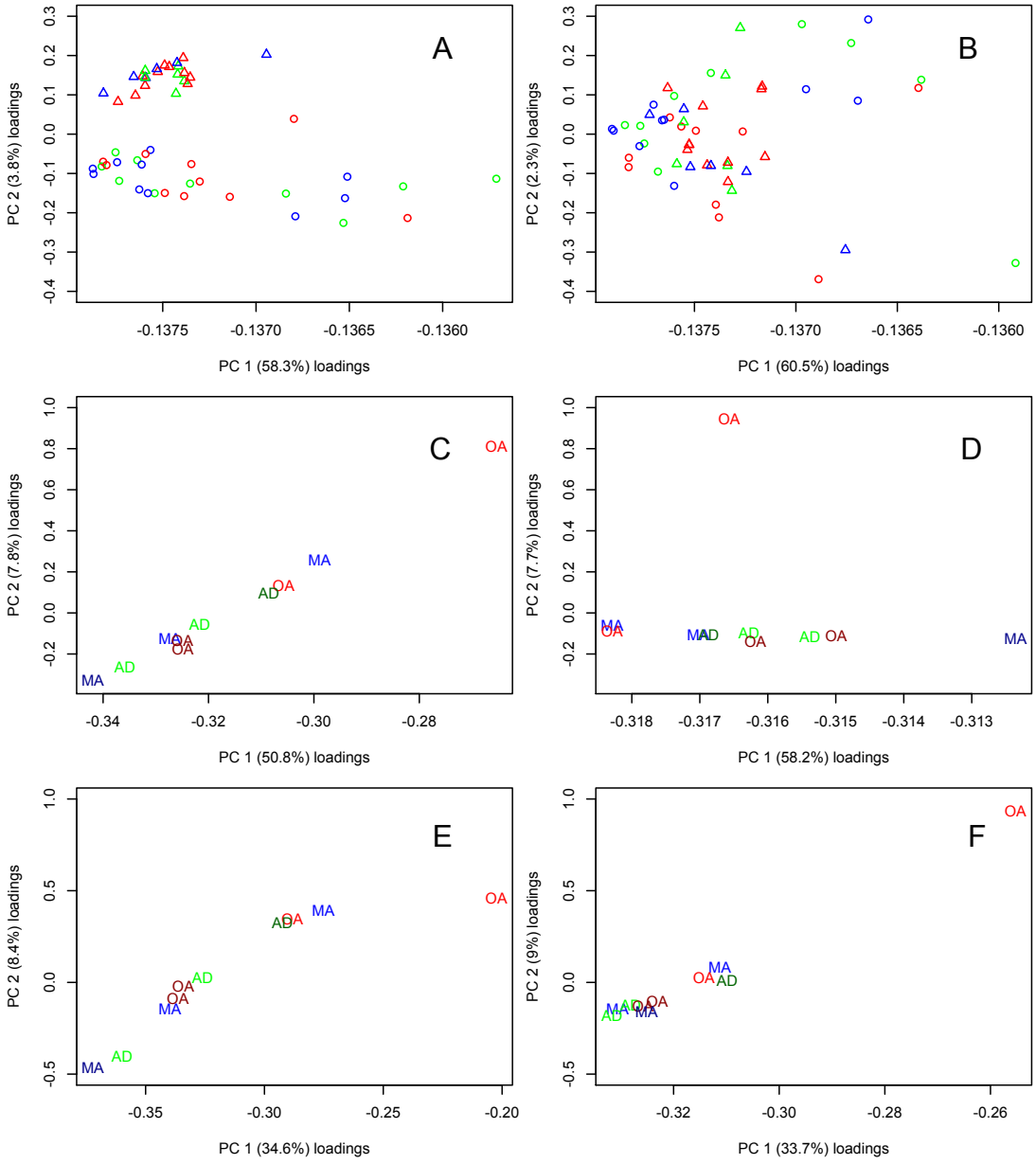


Figure 5.12: **A**: Loadings of the first two principal components estimated using array data from cohort 1 (circles) and cohort 2 (triangles) in which colours are the same as in Figure 5.8A, data were filtered by detection and quantile normalised. **B**: Similar to A, except data were batch-corrected by cohort using *ComBat*. **C**: Loadings of the first two principal components estimated using RNA-seq data from sample pools of cohort 1 (lighter colours) and cohort 2 (darker colours) based on gene-level *Maq* read-counts. **D**: Similar to C, except normalised gene-level RPKM expressions were used. **E**: Loadings of the first two principal components estimated using RNA-seq data from sample pools of cohort 1 (lighter colours) and cohort 2 (darker colours) based on exon-level *Maq* read-counts. **F**: Similar to E, except normalised exon-level RPKM expressions were used.

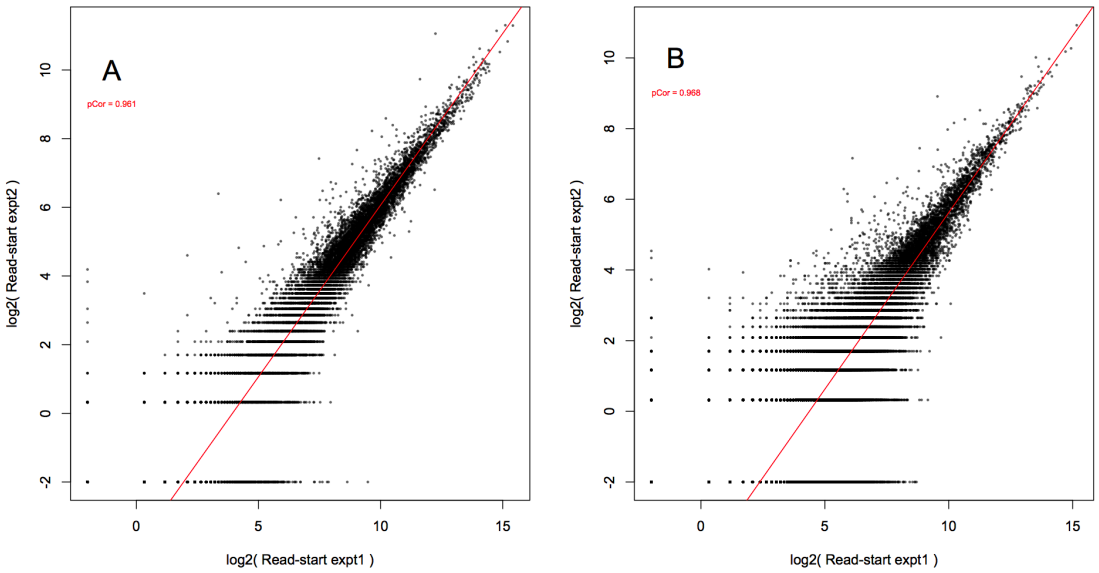


Figure 5.13: **A**: Gene-level expressions of sample 'AD_{p1}' as reported in the first (x-axis) and second (y-axis) RNA-seq experiment runs. *Casava* read-start counts were used to quantify the expression. Red line is of unit gradient, shifted to compensate for different library sizes between the samples. **B**: Same as **A**, except expressions are summarised at the exon-level.

annotated gene regions (Fig 5.14A). However the base-counting method generally leads to increased expression estimates at the exon-level and, in a large number of cases, leads to a much greater expression estimate compared to read-start position based counts (Fig 5.14B). This could be a result of short exons, for which the majority of a read lies within the exon, but it starts in a preceding exon and longer reads make this problem worse. It could also be due to incorrectly annotated exons in the reference, in which the annotation suggests the exon starts after it actually does in the sample transcriptome.

5.4 Discussion

5.4.1 Assessment of levels of variation in these RNA-seq data

The high level of correlation observed between technical the RNA-seq replicate sample-pairs at the gene- and exon-level agree with data obtained by others from the same core-facility [109, 111]. This was reflected in analyses testing the departure from random Poisson-noise between the replicates, in which a very small fraction of the total genes and exons showed lane-to-lane variation above that expected under the random model, but this was not the case at the probe-level.

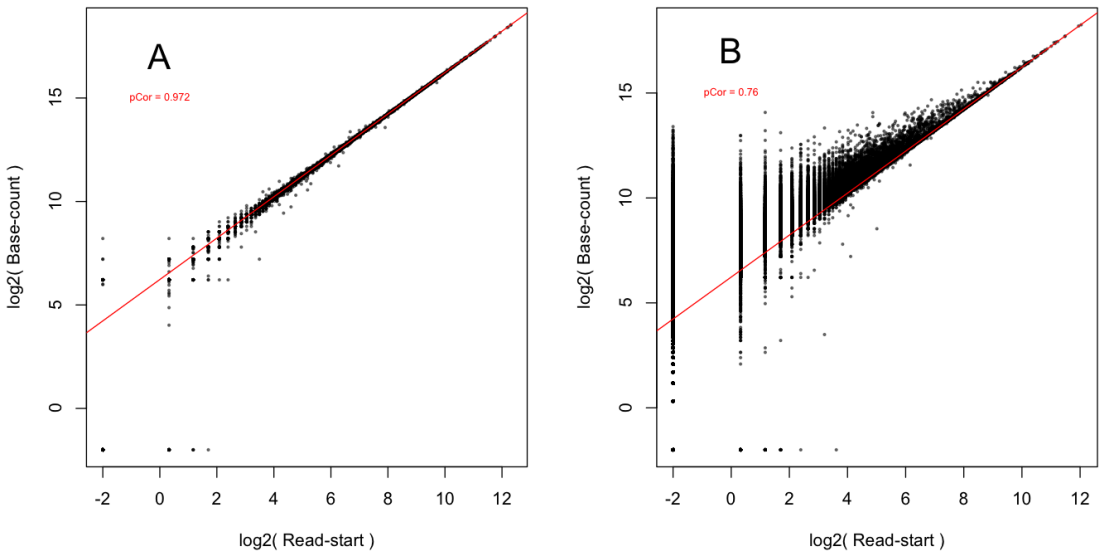


Figure 5.14: **A**: Gene-level RNA-seq expressions of sample 'AD_{p1}', from the second experiment, as reported by *Casava* read-start (x-axis) and read-bases (y-axis) quantification methods. Red line is unit gradient, shifted by $log_2(75)$ to compensate for the counting of bases rather than reads. **B**: Same as **A**, except expressions are summarised at the exon-level.

This was the first instance in which the technical replicates were compared at the array-probe level and the intention was to assess whether quantifying RNA-seq expression at such a low level would enable comparisons with microarray expressions without relying on the quality of the reference sequence annotation. It is not a surprise that the variability between technical replicates by RNA-seq at this resolution was high due to the much smaller window in which RNA-seq reads were mapped. This level of analysis is therefore of limited utility due to this technical variation in read-depth.

Compared to the very strong agreement between these technical replicates, run on the same chip and quantified using the same mapping software, the correlation between the same raw data analysed using different mapping software was weaker. Over a quarter of genes exhibited extra-Poisson variation between mapping methods, compared with 1% between technical replicates quantified using the same software. The main difference between the mapping methods, and the most likely explanation for the expression differences, is the use of trimmed reads prior to mapping and quantification with *Maq*. Therefore it is reasonable to posit that, especially when data quality is questionable, the choice of mapping and trimming reads to remove spurious bases is much more important than running on-chip technical replicate samples.

Correlation analyses are less suited to discrete count data, even after taking the

logarithm, than assessing departure from the random Poisson noise model for each feature. This is due to the effect of a large number of genes with no mapped reads inflating the correlation estimate over the dataset. The re-sequenced samples had poorer correlation to the original samples than that observed between mapping methods, however the fraction of genes with extra-Poisson variance was much lower, suggesting that even replicate library preparations (from the same RNA source) contribute less variability than would a nave choice of mapping method.

5.4.2 RNA-seq vs arrays

It is clear from analyses performed here, and those reported by others, that RNA-seq is more sensitive than microarrays at lower levels of expression. A large number of genes and exons that fall below the detection limit on the arrays were found to be highly expressed by RNA-seq, indicating the effect of the lower detection-limit of the arrays [87]. Indeed, the strongest agreement between the two technologies is at very high and very low levels of expression and, due to the different expression distributions output by each technology, direct data integration at the absolute expression level is unwise.

Unfortunately, the samples used in these experiments were found to exhibit high levels of biological variability in both the array and RNA-seq data and very few features remained significant following FDR adjustment. BeadChips consistently identified a greater number of probes for which the unadjusted p-values were reported to be significant, however this is likely due to the much larger number of biological replicates available on arrays. Both datasets are clearly underpowered to detect robust significant differences between the ages by pairwise comparisons, however ANOVA analyses on arrays using with all three age-groups were more successful at detecting changes in expression between them [207]. These analyses were not presented here as at the time of writing the RNA-seq analysis methods used were capable only of pairwise comparison and the goal was inter-technology comparison, rather than biological discovery.

It is likely that the pooling of samples prior to RNA-seq was the main cause of the low power. There are many factors, such as mild infections, that can affect gene-expression in the brain but may not present a strong phenotype. Pooling increases the chance that the cells or tissues under study contain abnormal cells from a diseased subject [221]. If such a subject is present in the pooled sample, it is impossible to detect it as an outlier and the increase in variability between biological replicates becomes much more difficult to cater for. It is likely that was the case in the pooled data obtained in these experiments.

Regardless of the effect of pooling, methods for identifying differential expression in these RNA-seq data were determined to be effective and that FDR adjustment

is not overly conservative in cases where there is clear evidence for a large number of differentially expressed genes. Following permutation of the six RNA-seq samples we suspected a potential error was made during sample preparation that caused one of the MA samples to be switched with an OA, due to the large number of genes found differentially expressed when the samples in lanes 2 and 3 were in the same pseudo age-group. However, after re-sequencing these samples it is clear that the biological variation between pooled samples was simply too high and the similarity of sample MA_{p1} and OA_{p1} had not been mislabeled. It is a shame that the samples were sequenced, twice, in pools rather than individually as this would have provided more data with which to assess the biological variability- but in the interests of cost and limited availability of starting material this was prohibited.

In the absence of robust, statistically significant results to compare between the technologies we used the next best thing. Comparison of reported fold-changes (FC) between the technologies illustrated a general positive trend. A large number of genes and exons were observed with large-FC by RNA-seq and low-FC by arrays, which were mostly populated by undetected features on the arrays. For the small number of genes and exons with a significant p-value on both RNA-seq and array the overall agreement between the technologies was reasonable. Fold-changes reported by arrays did not accurately reflect those reported by RNA-seq in terms of absolute change, however the predictive power was very strong when only the direction of change was considered.

Based on the good consistency between fold-changes, it is clear that differential expression between array and RNA-seq are equivalent and that a combination of high biological variability and pooled RNA-seq samples were the cause for low numbers of statistically robust results in these data. However, several studies have compared differential expression estimated by array to that from other methods, mostly qPCR, and discussed the compressed fold-changes obtained by the arrays. This also seems to be the case in our data; array fold-change estimates consistently lower, on average, than those derived from RNA-seq data.

The read-qualities for the re-sequenced samples in cohort 1 remained poor, but diagnostics indicated this was more likely due to degraded biological material rather than a technical fault on the sequencer. Also, perhaps due to a combination of poor-quality input RNA and multiplexed sequencing, the overall yield was much lower than in the data obtained from the first round of sequencing. However, a greater fraction of these new reads were successfully mapped with *Casava*, despite the overall poor read quality suggesting that the specific problem around base 35 in the old reads was largely responsible for the large number of discarded data. However, the correlation between reported read-counts in the old and the new sequenced samples was surprisingly high, but still around 7.5% of the genes and 1.3% of the exons exhibited variation that is

larger than that expected from random shot-noise.

As with the first dataset, PCA revealed large biological variation between the new cohort 1 samples at both the gene- and exon-level. Variation between the new cohort two samples appeared to be less, however the lack of any meaningful numbers of biological replicates in either cohort and small number of mapped reads make drawing firm conclusions from these data rather perilous.

5.4.3 Discussion of level of analysis and quantification method

It is clear from the comparison of the read-bases and read-start quantification methods within *Casava* that the former is much more accurate to the underlying biology and structure of the transcriptome than the latter. As expected, the two methods had little effect on expression estimates at the gene-level, but lead to a significant increase in the expressions estimated at the exon-level. Even the use of read-middle quantification, employed in our *Maq*-quantification pipeline, would be an improvement (at the exon-level) on the read-start method, allowing more flexibility in terms of the accuracy of the exon-locations as defined in the reference annotation.

Exon level comparison of RNA-seq with arrays (and within RNA-seq datasets) is perhaps more suitable than analysis at the gene level. Absolute quantification of RNA-seq technical replicates at the exon level is approximately equivalent, in terms of variability, than when reads are summarised over the length of the whole gene. Analyses at the level of individual exons is also more intuitive as these are the building blocks of protein-coding transcripts and subtleties can be lost when performing much less granular analyses at the gene-level. It is also possible to search for and quantify putative novel isoforms by careful analysis at the exon-level, which is not possible at the gene-level.

However, in terms of the relatively simplistic comparison of differential expression reported by the two technologies, the sacrifice of granularity for generality and ease of interpretation might favour gene-level analyses. Exon-level differential expression is hindered by the much larger number of required comparisons ($\sim$10 times as many exons as there are genes) potentially falling foul of perhaps overly conservative multiple testing corrections.

Analysis at the transcript level is the natural conclusion in the discussion of what biological unit of measurement to use in quantifying expression. Progress has been made in this approach and a number of software packages are available, including ERANGE [107], tophat/cufflinks [108, 215], Scripture [222], and IQSeq (unpublished: http://archive.gersteinlab.org/proj/rnaseq/IQSeq/). However it is worth bearing in mind that the assignment of reads, and indeed entire exons to individual transcripts

and the identification of novel isoforms is an extremely difficult procedure. Inclusion of known splice-sites has been reported to improve the identification of novel isoforms [108], presumably due to the reduction in search space following the initial mapping. Based on evidence between the different mapping/quantification approaches used in our data, this procedure is also highly prone to bias downstream analyses if care is not taken during the initial read QC and alignment.

5.5 Conclusion

Overall we found good agreement between raw expressions and fold-changes reported by RNA-seq and by arrays at both the exon- and the gene-level, however the different distributions of raw expressions preclude direct integration of these raw data. Integration of fold-changes reported by each technology is more promising, after compensating for the overestimation of changes by RNA-seq, although it is not unreasonable to assume that this agreement would increase on analysis of better-quality, and more comprehensive, RNA-seq data. It is likely that the high biological variability between mice, a low-powered study design, and pooling of samples prior to sequencing were the reasons for the failure to detect differential expression in these data. Many factors, such as mild infections, can affect gene-expression in the brain but may not present a strong phenotype and pooling increases the chance that cells or tissues under study contain abnormal cells from a diseased subject

Reflected in concerns about data quality, we found that the choice of method for mapping and quantifying the raw sequence reads can have a large effect on the expression estimates. Analyses at the gene- and exon-level were approximately equivalent in terms of the comparison with microarrays and also between technical & biological replicates on RNA-seq. However, transcript-level analysis is clearly the most desirable method, as it best reflects the underlying cellular processes, but this requires advanced mapping and quantification methods to provide reliable expressions; care must be taken to ensure this is performed correctly, or reported expressions and putative novel isoforms will be unreliable.

Chapter 6

General conclusions and future work

6.1 Sample preparation and choice of technical replicates

Development of technologies and methods for the study of molecular biology is approximately a punctuated equilibrium, in which short periods of rapid development are followed by much longer periods of refinement in which experiment data are gathered, analysed, and modelled. As part of the refinement process, the field learns and experiments with best practice to get the most out of the available technology to improve the scope, throughput, reliability, and accuracy of experimental methods. However, strange norms are sometimes adopted by the community that are believed to have a positive impact on accuracy or reliability of the obtained measurements, but in fact do neither.

A prime example of one such norm is in qPCR, where it is extremely common to use technical replicates of the qPCR step itself, typically in triplicate. Studies, including this one (Chapter 2), using qPCR technical replicates have shown the consistency of qPCR amplification and quantification to be highly reproducible [25] and it seems this convention is borne from a simple desire to insure against a failed reaction. However, in most instances we found that it is the process of sampling/extraction of nucleic material and/or the reverse-transcription step that introduce the most variability. Splitting and averaging technical replicates at higher levels, such as taking multiple tissue samples from the same subject, is therefore likely to improve the consistency of measurements between experimental subjects sharing a common phenotype and improve the statistical power of the experiment. Taking higher-level replicates obviously provides exactly the same protection against a failed reaction as the same number of reactions are run for each subject.

The convention to not use technical replicates in microarray experiments is again borne from a large number of studies comparing array expression measurements within [223, 188] and between [118, 182] platforms that have reported good overall agreement. However, as with qPCR, very few studies have investigated the effect of the various stages sample preparation on the reliability of reported expression levels and those that have reported poor consistency between different methods of amplification [73]. Chapter 3 demonstrated that batch effects can compromise the reproducibility of reportedly significantly differentially expressed genes between conditions, but that experiments with no technical replicates can still produce meaningful and reliable results if the experimental design is robust to such batches and appropriate batch-corrections are employed.

Several high-profile investigations such as MAQC [118] investigated inter-and intra-platform consistency of microarrays. However this study was limited to analysis of high-quality, commercial-grade reference RNAs and did not address vulnerability of the

cDNA-prep to sampling-variability. Chapter 4 sought to evaluate the relative effects of different sources of variation and demonstrated that the variability between biological replicate cultures and technical replicate RNA extractions is much greater than the technical noise introduced during downstream sample processing and array scanning. It is therefore a credit to the methods and to the technologies themselves that the experiment data they produce is generally so reliable. However this precision has the negative effect of entertaining many researchers' ignorance about biological and sampling variation and lack of understanding about reproducibility/accuracy leading to limited skepticism in their interpretation and reporting of results obtained from high-throughput approaches.

Systematic variation is common in all biological experiments [119] and can have a large influence on the interpretation of experiment data, so it is especially important to understand its impact in 'hypothesis generating' experiments such as microarray, RNA-seq, and, increasingly, qPCR. It is clear that such error is not always simply due to imprecision of the analysis technology, but also due to variation in the preparation of samples for analysis.

6.2 Variability of reported expressions and results

The large numbers of genes concurrently interrogated by high-throughput approaches coupled with often small numbers of independent biological observations result in studies that routinely have insufficient statistical power to confidently detect differential expression [224]. There has been some frustration in several fields on the perceived lack of reproducibility/overlap in gene-lists derived from different microarray analyses [225, 226, 227, 228], however there often exist technical and biological differences between studies that completely confound analyses and it is therefore not surprising that results are in poor agreement. *In-silico* differences can also play a large role in determining the reproducibility of results both between replicate studies, and re-analyses of identical data from the same study [228]. For example, poor documentation of statistical methods used has been cited as a common cause of poor reproducibility in re-analyses of published data [154]. A similar conclusion was reached by the MAQC following their stage-II analyses, in which identical data were provided to several different teams for statistical analysis and clustering, finding that- perhaps unsurprisingly -the proficiency of the statistician was a major factor determining the outcome of the analyses [229].

Sources of bias, especially that introduced by the different dates and times at which samples are processed further serve to threaten the reporting of accurate and reproducible results [230, 231]. The combined effect of all of these sources of biological,

technical, and computational variability is that thousands of biologically independent observations are required to produce robust and useful results on analysis of complex systems, especially in the clinical setting [232].

Effect of technical variation on reported expressions and results

In qPCR experiments it was shown that the effect of normalising a gene of interest (GOI) to a reference gene with a very different variance structure actually increased the overall variance for the GOI. This suggests the need to select a reference gene that is not affected by the experiment factor under investigation, but does co-vary with the GOI with respect to the variation introduced at each level of sample-prep. In terms of the magnitude of technical variation, most originates from sampling and in solid tissue up to 2.6-fold variation was observed between technical replicates. It is therefore fairly clear that such variation within and between experimental subjects has high potential to influence results of analyses for differential expression.

Data from replicate array hybridisations presented in chapter 3 showed a small systematic effect due to the day on which the arrays were processed. All the arrays were of perfectly reliable quality, with similarly excellent correlation ($> 97\%$) with those reported by MAQC. However, this small systematic variability had a large negative effect on the consistency between results of identical analyses for differential expression between technical replicate tumour samples. Previously published array analyses using technical replicate samples [182] reported that comparing gene-lists based on their reported statistical significances was less reliable than comparing them based on reported fold-changes between treatment groups. This is an unsatisfactory recommendation, but does support the observation that technical variation is the limiting factor for consistency of such differential expression tests, rather than some bulk error affecting the magnitude of the overall change between groups of samples.

A small number of studies have assessed technical replicates using RNA-seq and found extremely low noise in the reported expressions. However we found greater variation between different methods of quality control and mapping of the raw sequence reads to the reference. Although not strictly technical variation as defined throughout this thesis, differences in mapping illustrate the importance of performing careful computational analysis, especially during the initial expression quantification. Due to technical issues with the sequence data quality, it is not prudent to posit confident conclusions as to the origin or effect of technical variation observed in these data, nor from the comparison to the array data.

However, work to analyse the second set of sequenced data continues, although the sample size and the total number of reads are very small- limiting the usefulness and

power of such an analysis. It is also of interest to further explore the array probes that were found to express above background, but were not represented at either the gene- or exon-level by RNA-seq. An analysis of the compositional properties of the probes might reveal the extent to which they suffer from non-specific hybridisation, and an analysis of the mapping properties might reveal if some feature of the sample-prep and library-creation prior to RNA-seq was responsible for the inconsistency between the platforms.

6.3 Compensation for systematic technical variation

The two main considerations for obtaining a sound experiment design is where to take replicates and how to randomise/block samples such that systematic errors are not confounded with the biology. In terms of the former, we have shown that the relative error introduced in each of the various levels of sample-prep are generally predictable and will, in most cases, be smallest at the point of measurement and largest between experimental subjects or between replicate samples/RNA-extractions. There is the assumption that all genes are affected equally by each of these steps, however we and others have shown this to be a false assumption [119, 126, 127]. With the ever-decreasing cost of microarrays and technologies like Illumina's delivering multiple arrays on a single chip, it is reasonable that studies limited to the analysis of small numbers of subjects might benefit greatly from the use of technical replicates at the sampling level. Such replicates might improve statistical power by reducing within-group biological variability between the subjects. Also, in array experiments where the number of genes is very large, it is possible to use a method such as SVA [128] to diagnose the source(s) of the confounding noise. In cases where large numbers of observations are not available, such as qPCR, we have shown that relative contributions of the various levels can be estimated from pilot data- and even possible to automatically create optimal experiment designs.

6.4 In summary

It has been shown in chapters 3 and 4 that standard normalisation methods are not effective at reducing or removing systematic inter-run or inter-chip variation in microarray experiments and that more specific gene-level corrections are required to improve the correlation of technical replicate control and tumour samples. In chapter 2, reduction of technical variation in qPCR experiments can be achieved through targeting of replicates to the level at which most noise is introduced, following a pilot study in which the relative variation at each level can be estimated. Finally, in chapter 5, RNA-

seq technical replicates were found to have remarkably strong correlation, even between experiments performed many months apart, on different iterations of the technology, and where the output short-read sequences were of dubious quality. However, in the case of RNA-seq it is clear that the requirement of reliable and appropriate data analysis is more important than ever as different methods of quality control and assignment of the raw reads, as well as the method of quantification, can have a large impact on consistency of results. Variability between RNA-seq and microarrays was much larger than variability due to read-mapping, but correlations between reported direction and, to some extent, magnitude of differential expression between sample-groups were reasonably strong.

The analyses presented in this thesis have consistently reinforced the notion that high-throughput, whole-genome mRNA analysis methods are fully capable of producing highly reliable, reproducible, and accurate data. Systematic variation can threaten this remarkable reliability but with close collaboration, researchers and analysts can deal with these issues with relative speed and simplicity through careful and informed experiment planning and analysis. We, and others, have consistently shown that the incorporation of specific batch-correction methods should be a routine step in the analysis of these high-throughput data and that targeted experiments to determine the precise sources of non-biological variation will always be necessary to maximise the efficacy of such corrections. Greater confidence and accuracy of the hypotheses generated by these high-throughput techniques not only improves publishable results, but leads to more efficient allocation and use of resources for followup confirmatory analyses, either at the mRNA- or the protein-level.

Increasingly, as technology and methods mature, researchers are choosing to integrate many diverse sources of whole-genome data in an attempt both to better understand the wider biological context of their experiments and also to reduce the rate of false-positive results produced by each individual method. This is commonly achieved by finding and reporting only those features that are found to be significant in all, or a large fraction of, the experiments. Despite their effectiveness in reducing the frequency of false-positives, only increasing the accuracy and reliability of each individual experimental method can improve the typically poor power of such integrative analysis approaches. For example, new methods such as RNA-seq show great promise in aiding the analysis and interpretation of proteomics data derived from tandem mass-spectrometry; but to fully realise the power and potential that integrative, genome-wide mRNA, proteomic, and epigenetic analyses promises, it is critical that we ensure high confidence in the data and results obtained at each level.

Bibliography

[1] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, pp. 561–3, Aug 1970.

[2] T. J. Griffin, S. P. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood, and R. Aebersold, "Complementary profiling of gene expression at the transcriptome and proteome levels in saccharomyces cerevisiae," *Mol Cell Proteomics*, vol. 1, pp. 323–33, Apr 2002.

[3] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein, "Comparing protein abundance and mrna expression levels on a genomic scale," *Genome Biol*, vol. 4, p. 117, Jan 2003.

[4] D. A. Bitton, M. J. Okoniewski, Y. Connolly, and C. J. Miller, "Exon level integration of proteomics and microarray data," *BMC Bioinformatics*, vol. 9, p. 118, Jan 2008.

[5] D. A. Bitton, D. L. Smith, Y. Connolly, P. J. Scutt, and C. J. Miller, "An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome," *PLoS ONE*, vol. 5, p. e8949, Jan 2010.

[6] E. S. Lander, "Array of hope," *Nat Genet*, vol. 21, pp. 3–4, Jan 1999.

[7] J. C. Alwine, D. J. Kemp, and G. R. Stark, "Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes," *Proc Natl Acad Sci USA*, vol. 74, pp. 5350–4, Dec 1977.

[8] R. Higuchi, G. Dollinger, P. S. Walsh, and R. Griffith, "Simultaneous amplification and detection of specific dna sequences," *Biotechnology (NY)*, vol. 10, pp. 413–7, Apr 1992.

[9] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, pp. 467–70, Oct 1995.

[10] E. R. Mardis, "A decade's perspective on dna sequencing technology," *Nature*, vol. 470, pp. 198–203, Feb 2011.

[11] M. Metzker, "Sequencing technologies—the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2009.

[12] R. Tewhey, J. B. Warner, M. Nakano, B. Libby, M. Medkova, P. H. David, S. K. Kotsopoulos, M. L. Samuels, J. B. Hutchison, J. W. Larson, E. J. Topol, M. P. Weiner, O. Harismendy, J. Olson, D. R. Link, and K. A. Frazer, "Microdroplet-based pcr enrichment for large-scale targeted sequencing," *Nat Biotechnol*, vol. 27, pp. 1025–31, Nov 2009.

[13] A. Oshlack, M. D. Robinson, and M. D. Young, "From rna-seq reads to differential expression results," *Genome Biol*, vol. 11, p. 220, Jan 2010.

[14] A. Raj and A. van Oudenaarden, "Nature, nurture, or chance: stochastic gene expression and its consequences," *Cell*, vol. 135, pp. 216–26, Oct 2008.

[15] T. L. Fare, E. M. Coffey, H. Dai, Y. D. He, D. A. Kessler, K. A. Kilian, J. E. Koch, E. LeProust, M. J. Marton, M. R. Meyer, R. B. Stoughton, G. Y. Tokiwa, and Y. Wang, "Effects of atmospheric ozone on microarray data quality," *Anal Chem*, vol. 75, pp. 4672–5, Sep 2003.

[16] S. A. Bustin, "Absolute quantification of mrna using real-time reverse transcription polymerase chain reaction assays," *J Mol Endocrinol*, vol. 25, pp. 169–93, Oct 2000.

[17] M. Bengtsson, M. Hemberg, P. Rorsman, and A. Ståhlberg, "Quantification of mrna in single cells and modelling of rt-qpcr induced noise," *BMC Mol Biol*, vol. 9, p. 63, Jan 2008.

[18] C. Heid, J. Stevens, K. Livak, and P. Williams, "Real time quantitative pcr," *Genome Research*, vol. 6, pp. 986–994, Jan 1996.

[19] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim, "Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia," *Science*, vol. 230, pp. 1350–4, Dec 1985.

[20] A. Tichopad, M. Dilger, G. Schwarz, and M. W. Pfaffl, "Standardized determination of real-time pcr efficiency from a single reaction set-up," *Nucleic Acids Res*, vol. 31, p. e122, Oct 2003.

[21] A. M. Wang, M. V. Doyle, and D. F. Mark, "Quantitation of mrna by the polymerase chain reaction," *Proc Natl Acad Sci USA*, vol. 86, pp. 9717–21, Dec 1989.

[22] K. J. Livak, S. J. Flood, J. Marmaro, W. Giusti, and K. Deetz, "Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting pcr product and nucleic acid hybridization," *PCR Methods Appl*, vol. 4, pp. 357–62, Jun 1995.

[23] L. G. Lee, C. R. Connell, and W. Bloch, "Allelic discrimination by nick-translation pcr with fluorogenic probes," *Nucleic Acids Research*, vol. 21, pp. 3761–6, Aug 1993.

[24] M. W. Pfaffl, "A new mathematical model for relative quantification in real-time rt-pcr," *Nucleic Acids Research*, vol. 29, p. e45, May 2001.

[25] S. A. Bustin and T. Nolan, "Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction," *J Biomol Tech*, vol. 15, pp. 155–66, Sep 2004.

[26] V. Popovici, D. R. Goldstein, J. Antonov, R. Jaggi, M. Delorenzi, and P. Wirapati, "Selecting control genes for rt-qpcr using public microarray data," *BMC Bioinformatics*, vol. 10, p. 42, Jan 2009.

[27] O. Fedrigo, L. R. Warner, A. D. Pfefferle, C. C. Babbitt, P. Cruz-Gordillo, and G. A. Wray, "A pipeline to determine rt-qpcr control genes for evolutionary studies: application to primate gene expression across multiple tissues," *PLoS ONE*, vol. 5, Jan 2010.

[28] S. A. Bustin, V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, J. Vandesompele, and C. T. Wittwer, "The miqe guidelines: minimum information for publication of quantitative real-time pcr experiments," *Clinical Chemistry*, vol. 55, pp. 611–22, Apr 2009.

[29] S. A. Bustin, J.-F. Beaulieu, J. Huggett, R. Jaggi, F. S. B. Kibenge, P. A. Olsvik, L. C. Penning, and S. Toegel, "Miqe précis: Practical implementation of minimum standard guidelines for fluorescence-based quantitative real-time pcr experiments," *BMC Mol Biol*, vol. 11, p. 74, Jan 2010.

[30] S. A. Bustin, V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, J. Vandesompele, and C. T. Wittwer, "Primer sequence disclosure: A clarification of the miqe guidelines," *Clinical Chemistry*, Mar 2011.

[31] M. Baker, "qpcr: quicker and easier but don't be sloppy," *Nat Methods*, vol. 8, pp. 207–12, Mar 2011.

[32] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor, "Light-generated oligonucleotide arrays for rapid dna sequence analysis," *Proc Natl Acad Sci USA*, vol. 91, pp. 5022–6, May 1994.

[33] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart, "High density synthetic oligonucleotide arrays," *Nat Genet*, vol. 21, pp. 20–4, Jan 1999.

[34] K. Kuhn, "A novel, high-performance random array platform for quantitative gene expression profiling," *Genome Research*, vol. 14, pp. 2347–2356, Nov 2004.

[35] K. L. Gunderson, "Decoding randomly ordered dna arrays," *Genome Research*, vol. 14, pp. 870–877, May 2004.

[36] Y. Barash, E. Dehan, M. Krupsky, W. Franklin, M. Geraci, N. Friedman, and N. Kaminski, "Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays," *Bioinformatics*, vol. 20, pp. 839–46, Apr 2004.

[37] M. J. Dunning, M. L. Smith, M. E. Ritchie, and S. Tavare, "beadarray: R classes and methods for illumina bead-based data," *Bioinformatics*, vol. 23, pp. 2183–2184, Jun 2007.

[38] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed, "Comparison of methods for image analysis on cdna microarray data," *Journal of Computational and Graphical Statistics*, vol. 11, pp. 1–29, Jan 2002.

[39] M. Schena, R. A. Heller, T. P. Theriault, K. Konrad, E. Lachenmeier, and R. W. Davis, "Microarrays: biotechnology's discovery platform for functional genomics," *Trends Biotechnol*, vol. 16, pp. 301–6, Jul 1998.

[40] T. Okamoto, T. Suzuki, and N. Yamamoto, "Microarray fabrication with covalent attachment of dna using bubble jet technology," *Nat Biotechnol*, vol. 18, pp. 438–41, Apr 2000.

[41] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, vol. 30, p. e15, Feb 2002.

[42] G. Smyth, "Normalization of cdna microarray data," *Methods*, vol. 31, pp. 265–273, Dec 2003.

[43] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nat Biotechnol*, vol. 14, pp. 1675–80, Dec 1996.

[44] D. Baird, P. Johnstone, and T. Wilson, "Normalization of microarray data using a spatial mixed model analysis which includes splines," *Bioinformatics*, vol. 20, pp. 3196–205, Nov 2004.

[45] J.-B. Fan, M. S. Chee, and K. L. Gunderson, "Highly parallel genomic assays," *Nat Rev Genet*, vol. 7, pp. 632–44, Aug 2006.

[46] P. J. Gardina, T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams, and Y. Turpaz, "Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array," *BMC Genomics*, vol. 7, p. 325, Jan 2006.

[47] J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker, "Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays," *Science*, vol. 302, pp. 2141–4, Dec 2003.

[48] X. S. Liu, "Getting started in tiling microarray analysis," *PLoS Comput Biol*, vol. 3, pp. 1842–4, Oct 2007.

[49] M. J. Dunning, N. P. Thorne, I. Camilier, M. L. Smith, and S. Tavaré, "Quality control and low-level statistical analysis of illumina beadarrays," *REVSTAT*, vol. 4, no. 1, pp. 1–30, 2006.

[50] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of affymetrix genechip probe level data," *Nucleic Acids Research*, vol. 31, p. e15, Feb 2003.

[51] M. J. Dunning, N. L. Barbosa-Morais, A. G. Lynch, S. Tavare, and M. Ritchie, "Statistical issues in the analysis of illumina data," *BMC Bioinformatics*, vol. 9, p. 85, Jan 2008.

[52] Z. Wu and R. A. Irizarry, "Stochastic models inspired by hybridization theory for short oligonucleotide arrays," *J Comput Biol*, vol. 12, pp. 882–93, Jan 2005.

[53] G. A. Churchill, "Fundamentals of experimental design for cdna microarrays," *Nat Genet*, vol. 32 Suppl, pp. 490–5, Dec 2002.

[54] H. Auer, S. Lyianarachchi, D. Newsom, M. I. Klisovic, G. Marcucci, U. Marcucci, and K. Kornacker, "Chipping away at the chip bias: Rna degradation in microarray analysis," *Nat Genet*, vol. 35, pp. 292–3, Dec 2003.

[55] S. L. Berger and H. L. Cooper, "Very short-lived and stable mrnas from resting human lymphocytes," *Proc Natl Acad Sci USA*, vol. 72, pp. 3873–7, Oct 1975.

[56] J. N. McClintick and H. J. Edenberg, "Effects of filtering by present call on analysis of microarray experiments," *BMC Bioinformatics*, vol. 7, p. 49, Jan 2006.

[57] W. Talloen, D.-A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H. W. H. Göhlmann, "I/ni-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data," *Bioinformatics*, vol. 23, pp. 2897–902, Nov 2007.

[58] A. J. Hackstadt and A. M. Hess, "Filtering for increased power for microarray data analysis," *BMC Bioinformatics*, vol. 10, p. 11, Jan 2009.

[59] R. Bourgon, R. Gentleman, and W. Huber, "Independent filtering increases detection power for high-throughput experiments," *Proc Natl Acad Sci USA*, vol. 107, pp. 9546–51, May 2010.

[60] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249–64, Apr 2003.

[61] W. Huber, A. von Heydebreck, H. Sueltmann, A. Poustka, and M. Vingron, "Parameter estimation for the calibration and variance stabilization of microarray data," *Statistical applications in genetics and molecular biology*, vol. 2, p. Article3, Jan 2003.

[62] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe, "Model-based variance-stabilizing transformation for illumina microarray data," *Nucleic Acids Research*, vol. 36, p. e11, Jan 2008.

[63] W. Cleveland, S. Devlin, and E. Grosse, "Regression by local fitting: Methods, properties, and computational algorithms," *Journal of Econometrics*, Jan 1988.

[64] J. Quackenbush, "Microarray data normalization and transformation," *Nat Genet*, vol. 32, pp. 496–501, Dec 2002.

[65] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, pp. 185–93, Jan 2003.

[66] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, vol. 3, p. Article3, Jan 2004.

[67] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci USA*, vol. 98, pp. 5116–21, Apr 2001.

[68] J. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society. Series BStatistical Methodology*, vol. 64, pp. 479–498, Jan 2002.

[69] A. Hartemink, D. Gifford, T. Jaakkola, and R. Young, "Maximum likelihood estimation of optimal scaling factors for expression array normalization," *SPIE BIOS 2001*, Jan 2001.

[70] R. A. Verdugo, C. F. Deschepper, G. Muñoz, D. Pomp, and G. A. Churchill, "Importance of randomization in microarray experimental designs with illumina platforms," *Nucleic Acids Res*, vol. 37, pp. 5610–8, Sep 2009.

[71] M. Barnes, "Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms," *Nucleic Acids Research*, vol. 33, pp. 5914–5923, Oct 2005.

[72] J. P. Novak, M. C. Miller, and D. A. Bell, "Variation in fiberoptic bead-based oligonucleotide microarrays: dispersion characteristics among hybridization and biological replicate samples," *Biol Direct*, vol. 1, p. 18, Jan 2006.

[73] C. L. Wilson, S. D. Pepper, Y. Hey, and C. J. Miller, "Amplification protocols introduce systematic but reproducible errors into gene expression studies," *BioTechniques*, vol. 36, pp. 498–506, Mar 2004.

[74] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (miame)-toward standards for microarray data," *Nat Genet*, vol. 29, pp. 365–71, Dec 2001.

[75] I. H. G. S. Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931–45, Oct 2004.

[76] E. S. Lander, "Initial impact of the sequencing of the human genome," *Nature*, vol. 470, pp. 187–97, Feb 2011.

[77] F. Sanger, S. Nicklen, and A. R. Coulson, "Dna sequencing with chain-terminating inhibitors," *Proc Natl Acad Sci USA*, vol. 74, pp. 5463–7, Dec 1977.

[78] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in dna by primed synthesis with dna polymerase," *J Mol Biol*, vol. 94, pp. 441–8, May 1975.

[79] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church, "Advanced sequencing technologies: methods and goals," *Nature Reviews Genetics*, vol. 5, pp. 335–44, May 2004.

[80] E. R. Mardis, "Next-generation dna sequencing methods," *Annu Rev Genomics Hum Genet*, vol. 9, pp. 387–402, Jan 2008.

[81] R. A. Holt and S. J. M. Jones, "The new paradigm of flow cell sequencing," *Genome Res*, vol. 18, pp. 839–46, Jun 2008.

[82] E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies," *Genomics*, vol. 93, pp. 105–11, Feb 2009.

[83] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. Mccauley, P. Mcnitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. R. Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. Mccooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, pp. 53–9, Nov 2008.

[84] K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette,

S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T.-F. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Q. Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. Mccarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. Mcvean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. Mcewen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart, "A second generation human haplotype map of over 3.1 million snps," *Nature*, vol. 449, pp. 851–61, Oct 2007.

[85] P. J. Park, "Chip–seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics*, vol. 10, pp. 669–680, Aug 2009.

[86] P. W. Laird, "Principles and challenges of genome-wide dna methylation analysis," *Nature Reviews Genetics*, vol. 11, pp. 191–203, Feb 2010.

[87] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nat Rev Genet*, vol. 10, pp. 57–63, Jan 2009.

[88] R. D. Hawkins, G. C. Hon, and B. Ren, "Next-generation genomics: an integrative approach," *Nat Rev Genet*, vol. 11, pp. 476–486, Jun 2010.

[89] D. S. Gerhard, L. Wagner, E. A. Feingold, C. M. Shenmen, L. H. Grouse, G. Schuler, S. L. Klein, S. Old, R. Rasooly, P. Good, M. Guyer, A. M. Peck, J. G. Derge, D. Lipman, F. S. Collins, W. Jang, S. Sherry, M. Feolo, L. Misquitta, E. Lee, K. Rotmistrovsky, S. F. Greenhut, C. F. Schaefer, K. Buetow, T. I. Bonner, D. Haussler, J. Kent, M. Kiekhaus, T. Furey, M. Brent, C. Prange, K. Schreiber, N. Shapiro, N. K. Bhat, R. F. Hopkins, F. Hsie, T. Driscoll, M. B. Soares, T. L. Casavant, T. E. Scheetz, M. J. Brown-stein, T. B. Usdin, S. Toshiyuki, P. Carninci, Y. Piao, D. B. Dudekula, M. S. H. Ko, K. Kawakami, Y. Suzuki, S. Sugano, C. E. Gruber, M. R. Smith, B. Simmons, T. Moore, R. Waterman, S. L. Johnson, Y. Ruan, C. L. Wei, S. Mathavan, P. H. Gunaratne, J. Wu, A. M.

Garcia, S. W. Hulyk, E. Fuh, Y. Yuan, A. Sneed, C. Kowis, A. Hodgson, D. M. Muzny, J. McPherson, R. A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madari, A. C. Young, K. D. Wetherby, S. J. Granite, P. N. Kwong, C. P. Brinkley, R. L. Pearson, G. G. Bouffard, R. W. Blakesly, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. N. Butterfield, M. Griffith, O. L. Griffith, M. I. Krzywinski, N. Liao, R. Morin, R. Morrin, D. Palmquist, A. S. Petrescu, U. Skalska, D. E. Smailus, J. M. Stott, A. Schnerch, J. E. Schein, S. J. M. Jones, R. A. Holt, A. Baross, M. A. Marra, S. Clifton, K. A. Makowski, S. Bosak, J. Malek, and M. P. Team, "The status, quality, and expansion of the nih full-length cdna project: the mammalian gene collection (mgc)," *Genome Res*, vol. 14, pp. 2121–7, Oct 2004.

[90] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial analysis of gene expression," *Science*, vol. 270, pp. 484–7, Oct 1995.

[91] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran, "Gene expression analysis by massively parallel signature sequencing (mpss) on microbead arrays," *Nat Biotechnol*, vol. 18, pp. 630–4, Jun 2000.

[92] F. Liu, T.-K. Jenssen, J. Trimarchi, C. Punzo, C. L. Cepko, L. Ohno-Machado, E. Hovig, and W. P. Kuo, "Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates," *BMC Genomics*, vol. 8, p. 153, Jan 2007.

[93] J. Chen, V. Agrawal, M. Rattray, M. A. L. West, D. A. S. Clair, R. W. Michelmore, S. J. Coughlan, and B. C. Meyers, "A comparison of microarray and mpss technology platforms for expression analysis of arabidopsis," *BMC Genomics*, vol. 8, p. 414, Jan 2007.

[94] J. Chen and M. Rattray, "Analysis of tag-position bias in mpss technology," *BMC Genomics*, vol. 7, p. 77, Jan 2006.

[95] R. Brandenberger, I. Khrebtukova, R. S. Thies, T. Miura, C. Jingli, R. Puri, T. Vasicek, J. Lebkowski, and M. Rao, "Mpss profiling of human embryonic stem cells," *BMC Dev Biol*, vol. 4, p. 10, Aug 2004.

[96] O. Morozova, M. Hirst, and M. A. Marra, "Applications of new sequencing technologies for transcriptome analysis," *Annu Rev Genomics Hum Genet*, vol. 10, pp. 135–51, Jan 2009.

[97] S. Marguerat and J. Bähler, "Rna-seq: from technology to biology," *Cell. Mol. Life Sci.*, vol. 67, pp. 569–79, Feb 2010.

[98] I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner, "Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+c)-biased genomes," *Nat Methods*, vol. 6, pp. 291–5, Apr 2009.

[99] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Res*, vol. 38, p. e131, Jul 2010.

[100] L. Mamanova, R. M. Andrews, K. D. James, E. M. Sheridan, P. D. Ellis, C. F. Langford, T. W. B. Ost, J. E. Collins, and D. J. Turner, "Frt-seq: amplification-free, strand-specific transcriptome sequencing," *Nat Meth*, vol. 7, pp. 130–2, Feb 2010.

[101] D. J. Sugarbaker, W. G. Richards, G. J. Gordon, L. Dong, A. D. Rienzo, G. Maulik, J. N. Glickman, L. R. Chirieac, M.-L. Hartman, B. E. Taillon, L. Du, P. Bouffard, S. F. Kingsmore, N. A. Miller, A. D. Farmer, R. V. Jensen, S. R. Gullans, and R. Bueno,

"Transcriptome sequencing of malignant pleural mesothelioma tumors," *Proc Natl Acad Sci USA*, vol. 105, pp. 3521–6, Mar 2008.

[102] T. T. Perkins, R. A. Kingsley, M. C. Fookes, P. P. Gardner, K. D. James, L. Yu, S. A. Assefa, M. He, N. J. Croucher, D. J. Pickard, D. J. Maskell, J. Parkhill, J. Choudhary, N. R. Thomson, and G. Dougan, "A strand-specific rna-seq analysis of the transcriptome of the typhoid bacillus salmonella typhi," *PLoS Genet*, vol. 5, p. e1000569, Jul 2009.

[103] S. Pepke, B. Wold, and A. Mortazavi, "Computation for chip-seq and rna-seq studies," *Nat Meth*, vol. 6, pp. S22–32, Nov 2009.

[104] P. Flicek and E. Birney, "Sense from sequence reads: methods for alignment and assembly," *Nat Meth*, vol. 6, pp. S6–S12, Nov 2009.

[105] B. J. Haas and M. C. Zody, "Advancing rna-seq analysis," *Nat Biotechnol*, vol. 28, pp. 421–3, May 2010.

[106] H. Li, J. Ruan, and R. Durbin, "Mapping short dna sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, pp. 1851–8, Nov 2008.

[107] A. Mortazavi, B. A. Williams, K. Mccue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nat Meth*, vol. 5, pp. 621–8, Jul 2008.

[108] C. Trapnell, L. Pachter, and S. L. Salzberg, "Tophat: discovering splice junctions with rna-seq," *Bioinformatics*, vol. 25, pp. 1105–11, May 2009.

[109] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res*, vol. 18, pp. 1509–17, Sep 2008.

[110] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments," *BMC Bioinformatics*, vol. 11, p. 94, Jan 2010.

[111] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad, "Sex-specific and lineage-specific alternative splicing in primates," *Genome Research*, vol. 20, pp. 180–9, Feb 2010.

[112] A. Oshlack and M. J. Wakefield, "Transcript length bias in rna-seq data confounds systems biology," *Biol Direct*, vol. 4, p. 14, Jan 2009.

[113] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "Degseq: an r package for identifying differentially expressed genes from rna-seq data," *Bioinformatics*, vol. 26, pp. 136–8, Jan 2010.

[114] M. D. Robinson and G. K. Smyth, "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics*, vol. 23, pp. 2881–7, Nov 2007.

[115] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by rna sequencing," *Science*, vol. 320, pp. 1344–9, Jun 2008.

[116] W. Huber and S. Anders, "Differential expression analysis for sequence count data," *Nature Precedings*, pp. 1–15, May 2010.

[117] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139–40, Jan 2010.

[118] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. D. Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. Leclerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. Mcdaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker, "The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nat Biotechnol*, vol. 24, pp. 1151–61, Sep 2006.

[119] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature reviews Genetics*, vol. 11, pp. 733–739, Oct 2010.

[120] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan, "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression," *Proc Natl Acad Sci USA*, vol. 101, pp. 9309–14, Jun 2004.

[121] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, pp. 572–7, Feb 2002.

[122] R. S. Spielman, L. A. Bastone, J. T. Burdick, M. Morley, W. J. Ewens, and V. G. Cheung, "Common genetic variants account for differences in gene expression among ethnic groups," *Nat Genet*, vol. 39, pp. 226–31, Feb 2007.

[123] J. M. Akey, S. Biswas, J. T. Leek, and J. D. Storey, "On the design and analysis of gene expression studies in human populations," *Nat Genet*, vol. 39, pp. 807–8; author reply 808–9, Jul 2007.

[124] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 559–572, Jun 1901.

[125] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[126] A. H. Sims, G. J. Smethurst, Y. Hey, M. J. Okoniewski, S. D. Pepper, A. Howell, C. J. Miller, and R. B. Clarke, "The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis," *BMC medical genomics*, vol. 1, p. 42, Jan 2008.

[127] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics (Oxford, England)*, vol. 8, pp. 118–27, Jan 2007. ComBat.

[128] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genet*, vol. 3, pp. 1724–35, Sep 2007.

[129] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc Natl Acad Sci USA*, vol. 97, pp. 10101–6, Aug 2000.

[130] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron, "Adjustment of systematic microarray data biases," *Bioinformatics*, vol. 20, pp. 105–14, Jan 2004.

[131] K. Owzar, W. T. Barry, S.-H. Jung, I. Sohn, and S. L. George, "Statistical challenges in preprocessing in microarray experiments in cancer," *Clin Cancer Res*, vol. 14, pp. 5959–66, Oct 2008.

[132] M. Kubista and N. Zoric, "The real-time polymerase chain reaction," *Molecular Aspects of Medicine*, pp. 95–125, Apr 2006.

[133] T. Nolan, R. E. Hands, and S. A. Bustin, "Quantification of mrna using real-time rt-pcr," *Nature Protocol*, vol. 1, pp. 1559–1582, Jan 2006.

[134] M. L. Wong and J. F. Medrano, "Real-time pcr for mrna quantitation," *BioTechniques*, vol. 39, pp. 75–85, Jul 2005.

[135] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative pcr and the 2(-delta delta c(t)) method," *Methods*, vol. 25, pp. 402–8, Dec 2001.

[136] R. A. Fisher, "The design of experiments," p. 248, Jan 1971.

[137] J. Cohen, "Statistical power analysis for the behavioral sciences," p. 567, Jan 1988.

[138] P. R. Rosenbaum, "Replicating effects and biases," *Amer. Statist.*, vol. 55, no. 3, pp. 223–227, 2001.

[139] R. Lenth, "Some practical guidelines for effective sample size determination," *The American Statistician*, vol. 55, pp. 187–193, Aug 2001.

[140] L. Thomas, "Retrospective power analysis," *Conservation Biology*, vol. 11, pp. 276–280, Feb 1997.

[141] A. Ståhlberg, J. Håkansson, X. Xian, H. Semb, and M. Kubista, "Properties of the reverse transcription reaction in mrna quantification," *Clinical Chemistry*, vol. 50, pp. 509–15, Mar 2004.

[142] E. Limpert, W. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues," *Bioscience*, vol. 51, pp. 341–352, Jan 2001.

[143] A. L. Oberg and D. W. Mahoney, "Linear mixed effects models," *Methods Mol Biol*, vol. 404, pp. 213–34, Jan 2007.

[144] A. Tichopad, R. Kitchen, I. Riedmaier, C. Becker, A. Ståhlberg, and M. Kubista, "Design and optimization of reverse-transcription quantitative pcr experiments," *Clinical Chemistry*, vol. 55, pp. 1816–23, Oct 2009.

[145] G. Snedecor and W. Cochran, *Statistical Methods (8th edition)*. 1989.

[146] J. Logan, K. Edwards, and N. Saunders, "Real-time pcr: Current technology and applications," p. 284, Feb 2009.

[147] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[148] MultiDAnalyses, "Genex software," 2009.

[149] A. Ståhlberg, M. Kubista, and M. Pfaffl, "Comparison of reverse transcriptases in gene expression analysis," *Clinical Chemistry*, vol. 50, pp. 1678–80, Sep 2004.

[150] A. H. Sims, "Bioinformatics and breast cancer: what can high-throughput genomic approaches actually tell us?," *Journal of Clinical Pathology*, vol. 62, pp. 879–885, Oct 2009.

[151] S. Ramaswamy and T. R. Golub, "Dna microarrays in clinical oncology," *J Clin Oncol*, vol. 20, pp. 1932–41, Apr 2002.

[152] R. Clarke, H. Ressom, A. Wang, J. Xuan, M. Liu, E. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nat Rev Cancer*, vol. 8, no. 1, pp. 37 – 49, 2008.

[153] K. A. Baggerly and K. R. Coombes, "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology," *Annals of Applied Statistics*, vol. 3, pp. 1–26, Sep 2009.

[154] J. P. A. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort, "Repeatability of published microarray gene expression analyses," *Nat Genet*, vol. 41, pp. 149–55, Feb 2009.

[155] T. Chu, S. Deng, R. Wolfinger, R. Paules, and H. Hamadeh, "Cross-site comparison of gene expression data reveals high similarity," *Environ Health Perspect*, vol. 112, no. 4, pp. 449 – 455, 2004.

[156] K. Thompson and P. Pine, "Comparison of the diagnostic performance of human whole genome microarrays using mixed-tissue rna reference samples," *Toxicol Lett*, vol. 186, no. 1, pp. 58 – 61, 2009.

[157] P. K. Tan, T. J. Downey, E. L. Spitznagel, P. Xu, D. Fu, D. S. Dimitrov, R. A. Lempicki, B. M. Raaka, and M. C. Cam, "Evaluation of gene expression measurements from commercial microarray platforms," *Nucleic Acids Research*, vol. 31, pp. 5676–84, Oct 2003.

[158] D. Eggle, S. Debey-Pascher, M. Beyer, and J. Schultze, "The development of a comparison approach for illumina bead chips unravels unexpected challenges applying newest generation microarrays," *BMC Bioinformatics*, vol. 10, p. 186, 2009.

[159] K. Baggerly, K. Coombes, and E. Neeley, "Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer," *J Clin Oncol*, vol. 26, no. 7, pp. 1186 – 1187, 2008.

[160] W. Shi, A. Banerjee, M. E. Ritchie, S. Gerondakis, and G. K. Smyth, "Illumina wg-6 beadchip strips should be normalized separately," *BMC Bioinformatics*, vol. 10, p. 372, Jan 2009.

[161] D. F. Ransohoff, "Bias as a threat to the validity of cancer molecular-marker research," *Nat Rev Cancer*, vol. 5, pp. 142–9, Feb 2005.

[162] D. Ransohoff and M. Gourlay, "Sources of bias in specimens for research about molecular markers for cancer," *J Clin Oncol*, vol. 28, no. 4, pp. 698 – 704, 2010.

[163] W. L. Walker, I. H. Liao, D. L. Gilbert, B. Wong, K. S. Pollard, C. E. Mcculloch, L. Lit, and F. R. Sharp, "Empirical bayes accomodation of batch-effects in microarray data using identical replicate reference samples: application to rna expression profiling of blood from duchenne muscular dystrophy patients," *BMC Genomics*, vol. 9, p. 494, Jan 2008.

[164] K. Thompson, P. Pine, B. Rosenzweig, Y. Turpaz, and J. Retief, "Characterization of the effect of sample quality on high density oligonucleotide microarray data using progressively degraded rat liver rna," *BMC Biotechnol*, vol. 7, p. 57, 2007.

[165] C. Acharya, D. Hsu, C. Anders, A. Anguiano, K. Salter, K. Walters, R. Redman, S. Tuchman, C. Moylan, S. Mukherjee, W. Barry, H. Dressman, G. Ginsburg, K. Marcom, K. Garman, G. Lyman, J. Nevins, and A. Potti, "Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer," *Jama*, vol. 299, no. 13, pp. 1574 – 1587, 2008.

[166] I. Ben-Porath, M. Thomson, V. Carey, R. Ge, G. Bell, A. Regev, and R. Weinberg, "An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors," *Nat Genet*, vol. 40, no. 5, pp. 499 – 507, 2008.

[167] Z. Zhang, D. Chen, and D. A. Fenstermacher, "Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome," *BMC Genomics*, vol. 8, p. 331, Jan 2007.

[168] R. Shen, D. Ghosh, and A. Chinnaiyan, "Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data," *BMC Genomics*, vol. 5, no. 1, p. 94, 2004.

[169] A. Teschendorff, A. Naderi, N. Barbosa-Morais, S. Pinder, I. Ellis, S. Aparicio, J. Brenton, and C. Caldas, "A consensus prognostic gene expression classifier for er positive breast cancer," *Genome Biol*, vol. 7, no. 10, p. R101, 2006.

[170] V. S. Sabine, A. H. Sims, E. J. Macaskill, L. Renshaw, J. S. Thomas, J. M. Dixon, and J. M. S. Bartlett, "Gene expression profiling of response to mtor inhibitor everolimus in pre-operatively treated post-menopausal women with oestrogen receptor-positive breast cancer," *Breast Cancer Res Treat*, vol. 122, pp. 419–28, Jul 2010.

[171] N. Barbosa-Morais, M. Dunning, S. Samarajiwa, J. Darot, M. Ritchie, A. Lynch, and S. Tavaré, "A re-annotation pipeline for illumina beadarrays: improving the interpretation of gene expression data," *Nucleic Acids Research*, Nov 2009.

[172] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A.-L. Børresen-Dale, and D. Botstein, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proc Natl Acad Sci USA*, vol. 100, pp. 8418–23, Jul 2003.

[173] M. Mullins, L. Perreard, J. Quackenbush, N. Gauthier, S. Bayer, M. Ellis, J. Parker, C. Perou, A. Szabo, and P. Bernard, "Agreement in breast cancer classification between microarray and quantitative reverse transcription pcr from fresh-frozen and formalin-fixed, paraffin-embedded tissues," *Clin Chem*, vol. 53, no. 7, pp. 1273 – 1279, 2007.

[174] K. Thompson, B. Rosenzweig, P. Pine, J. Retief, Y. Turpaz, C. Afshari, H. Hamadeh, M. Damore, M. Boedigheimer, E. Blomme, R. Ciurlionis, J. Waring, J. Fuscoe, R. Paules, C. Tucker, T. Fare, E. Coffey, Y. He, P. Collins, K. Jarnagin, S. Fujimoto, B. Ganter, G. Kiser, T. Kaysser-Kranich, J. Sina, and F. Sistare, "Use of a mixed tissue rna design

for performance assessments on multiple microarray formats," *Nucleic Acids Res*, vol. 33, no. 22, p. e187, 2005.

[175] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biol*, vol. 5, p. R80, Jan 2004.

[176] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, pp. 299 – 314, 1996.

[177] G. K. Smyth, J. Michaud, and H. S. Scott, "Use of within-array replicate spots for assessing differential expression in microarray experiments," *Bioinformatics*, vol. 21, pp. 2067–75, May 2005.

[178] G. LANCE and W. WILLIAMS, "A general theory of classificatory sorting strategies," *The Computer Journal*, vol. 9, no. 4, pp. 373 – 380, 1967.

[179] W. P. Kuo, T.-K. Jenssen, A. J. Butte, L. Ohno-Machado, and I. S. Kohane, "Analysis of matched mrna measurements from two different microarray technologies," *Bioinformatics*, vol. 18, pp. 405–12, Mar 2002.

[180] A. H. Sims, K. R. Ong, R. B. Clarke, and A. Howell, "High-throughput genomic technology in research and clinical management of breast cancer. exploiting the potential of gene expression profiling: is it ready for the clinic?," *Breast Cancer Res*, vol. 8, p. 214, Jan 2006.

[181] W. Jin, R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson, "The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster," *Nat Genet*, vol. 29, pp. 389–95, Dec 2001.

[182] L. Guo, E. K. Lobenhofer, C. Wang, R. Shippy, S. C. Harris, L. Zhang, N. Mei, T. Chen, D. Herman, F. M. Goodsaid, P. Hurban, K. L. Phillips, J. Xu, X. Deng, Y. A. Sun, W. Tong, Y. P. Dragan, and L. Shi, "Rat toxicogenomic study reveals analytical consistency across microarray platforms," *Nat Biotechnol*, vol. 24, pp. 1162–9, Sep 2006.

[183] L. Shi, W. D. Jones, R. V. Jensen, S. C. Harris, R. G. Perkins, F. M. Goodsaid, L. Guo, L. J. Croner, C. Boysen, H. Fang, F. Qian, S. Amur, W. Bao, C. C. Barbacioru, V. Bertholet, X. M. Cao, T.-M. Chu, P. J. Collins, X.-H. Fan, F. W. Frueh, J. C. Fuscoe, X. Guo, J. Han, D. Herman, H. Hong, E. S. Kawasaki, Q.-Z. Li, Y. Luo, Y. Ma, N. Mei, R. L. Peterson, R. K. Puri, R. Shippy, Z. Su, Y. A. Sun, H. Sun, B. Thorn, Y. Turpaz, C. Wang, S. J. Wang, J. A. Warrington, J. C. Willey, J. Wu, Q. Xie, L. Zhang, L. Zhang, S. Zhong, R. D. Wolfinger, and W. Tong, "The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies," *BMC Bioinformatics*, vol. 9 Suppl 9, p. S10, Jan 2008.

[184] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, pp. 207–10, Jan 2002.

[185] A. Brazma, M. Kapushesky, H. Parkinson, U. Sarkans, and M. Shojatalab, "Data storage and analysis in arrayexpress," *Meth Enzymol*, vol. 411, pp. 370–86, Jan 2006.

[186] R. R. Kitchen, V. S. Sabine, A. H. Sims, E. J. Macaskill, L. Renshaw, J. S. Thomas, J. I. van Hemert, J. M. Dixon, and J. M. S. Bartlett, "Correcting for intra-experiment variation in illumina beadchip data is necessary to generate robust gene-expression profiles," *BMC Genomics*, vol. 11, p. 134, Feb 2010.

[187] D. van der Veen, J. M. Oliveira, W. A. M. van den Berg, and L. H. de Graaff, "Analysis of variance components reveals the contribution of sample processing to transcript variation," *Appl Environ Microbiol*, vol. 75, pp. 2414–22, Apr 2009.

[188] W. Shi, A. Oshlack, and G. K. Smyth, "Optimizing the noise versus bias trade-off for illumina whole genome expression beadchips," *Nucleic acids research*, Oct 2010.

[189] S. O. Zakharkin, K. Kim, T. Mehta, L. Chen, S. Barnes, K. E. Scheirer, R. S. Parrish, D. B. Allison, and G. P. Page, "Sources of variation in affymetrix microarray experiments," *BMC Bioinformatics*, vol. 6, p. 214, Jan 2005.

[190] A. E. Pozhitkov, D. Tautz, and P. A. Noble, "Oligonucleotide microarrays: widely applied–poorly understood," *Brief Funct Genomic Proteomic*, vol. 6, pp. 141–8, Jun 2007.

[191] R. Owczarzy, P. M. Vallone, F. J. Gallo, T. M. Paner, M. J. Lane, and A. S. Benight, "Predicting sequence-dependent melting stability of short duplex dna oligomers," *Biopolymers*, vol. 44, pp. 217–39, Jan 1997.

[192] J. Neter, W. Wasserman, and M. H. Kutner, *Applied linear statistical models: regression, analysis of variance, and ....* Jan 1985.

[193] R. R. Kitchen, M. Kubista, and A. Tichopad, "Statistical aspects of quantitative real-time pcr experiment design," *Methods (San Diego, Calif)*, vol. 50, pp. 231–6, Apr 2010.

[194] J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar, "Linear and nonlinear mixed effects models," *R package version*, vol. 3, pp. 1–65, 2005.

[195] M. Lindstrom and D. Bates, "Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data," *Journal of the American Statistical Association*, vol. 83, pp. 1014–1022, Dec 1988.

[196] N. Laird and J. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, pp. 963–974, Dec 1982.

[197] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," *Genome Biol*, vol. 10, p. R25, Jan 2009.

[198] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott, "Ncbi reference sequences: current status, policy and new initiatives," *Nucleic acids research*, vol. 37, pp. D32–6, Jan 2009.

[199] R. Herwig, A. O. Schmitt, M. Steinfath, J. O'Brien, H. Seidel, S. Meier-Ewert, H. Lehrach, and U. Radelof, "Information theoretical probe selection for hybridisation experiments," *Bioinformatics*, vol. 16, pp. 890–8, Oct 2000.

[200] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput dna sequencing," *Nucleic Acids Research*, vol. 36, p. e105, Sep 2008.

[201] L. Mittempergher, J. J. D. Ronde, M. Nieuwland, R. M. Kerkhoven, I. Simon, E. J. T. Rutgers, L. F. A. Wessels, and L. J. V. Veer, "Gene expression profiles from formalin fixed paraffin embedded breast cancer tissue are largely comparable to fresh frozen matched tissue," *PLoS ONE*, vol. 6, p. e17163, Jan 2011.

[202] W. Tong, A. B. Lucas, R. Shippy, X. Fan, H. Fang, H. Hong, M. S. Orr, T.-M. Chu, X. Guo, P. J. Collins, Y. A. Sun, S.-J. Wang, W. Bao, R. D. Wolfinger, S. Shchegrova, L. Guo, J. A. Warrington, and L. Shi, "Evaluation of external rna controls for the assessment of microarray performance," *Nat Biotechnol*, vol. 24, pp. 1132–9, Sep 2006.

[203] R. D. Canales, Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, R. R. Samaha, L. Shi, W. Yang, L. Zhang, and F. M. Goodsaid, "Evaluation of dna microarray results with quantitative gene expression platforms," *Nat Biotechnol*, vol. 24, pp. 1115–22, Sep 2006.

[204] N. Novoradovskaya, M. L. Whitfield, L. S. Basehore, A. Novoradovsky, R. Pesich, J. Usary, M. Karaca, W. K. Wong, O. Aprelikova, M. Fero, C. M. Perou, D. Botstein, and J. Braman, "Universal reference rna as a standard for microarray experiments," *BMC Genomics*, vol. 5, p. 20, Mar 2004.

[205] M. A. Mongan, A. Mongan, M. Higgins, P. S. Pine, S. Pine, C. Afshari, and H. Hamadeh, "Assessment of repeated microarray experiments using mixed tissue rna reference samples," *BioTechniques*, vol. 45, pp. 283–92, Sep 2008.

[206] K. L. Thompson, B. A. Rosenzweig, P. S. Pine, J. Retief, Y. Turpaz, C. A. Afshari, H. K. Hamadeh, M. A. Damore, M. Boedigheimer, E. Blomme, R. Ciurlionis, J. F. Waring, J. C. Fuscoe, R. Paules, C. J. Tucker, T. Fare, E. M. Coffey, Y. He, P. J. Collins, K. Jarnagin, S. Fujimoto, B. Ganter, G. Kiser, T. Kaysser-Kranich, J. Sina, and F. D. Sistare, "Use of a mixed tissue rna design for performance assessments on multiple microarray formats," *Nucleic acids research*, vol. 33, p. e187, Jan 2005.

[207] K. A. Bordner, R. R. Kitchen, B. Carlyle, E. D. George, M. C. Mahajan, S. M. Mane, J. R. Taylor, and A. A. Simen, "Parallel declines in cognition, motivation, and locomotion in aging mice: association with immune gene upregulation in the medial prefrontal cortex," *Experimental gerontology*, Mar 2011.

[208] K. A. Bordner, E. Au, B. C. Carlyle, A. Duque, R. R. Kitchen, T. Lam, C. Colangelo, K. Stone, T. Abbott, S. M. Mane, A. Nairn, and A. A. Simen, "Functional genomic and proteomic analysis reveals disruption of myelin-related genes and translation in a mouse model of early life neglect," *Frontiers in Neurogenomics*, (in press).

[209] M. J. Okoniewski and C. J. Miller, "Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations," *BMC Bioinformatics*, vol. 7, p. 276, Jan 2006.

[210] T. E. Royce, J. S. Rozowsky, and M. B. Gerstein, "Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification," *Nucleic acids research*, vol. 35, p. e99, Jan 2007.

[211] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder, "What is a gene, post-encode? history and updated definition," *Genome Res*, vol. 17, pp. 669–81, Jun 2007.

[212] T. A. Clark, C. W. Sugnet, and M. Ares, "Genomewide analysis of mrna processing in yeast using splicing-specific microarrays," *Science*, vol. 296, pp. 907–10, May 2002.

[213] D. Zheng, A. Frankish, R. Baertsch, P. Kapranov, A. Reymond, S. W. Choo, Y. Lu, F. Denoeud, S. E. Antonarakis, M. Snyder, Y. Ruan, C.-L. Wei, T. R. Gingeras, R. Guigó, J. Harrow, and M. B. Gerstein, "Pseudogenes in the encode regions: consensus annotation, analysis of transcription, and evolution," *Genome Research*, vol. 17, pp. 839–51, Jun 2007.

[214] B. T. Wilhelm, S. Marguerat, I. Goodhead, and J. Bähler, "Defining transcribed regions using rna-seq," *Nature Protocols*, vol. 5, pp. 255–66, Jan 2010.

[215] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nat Biotechnol*, vol. 28, pp. 511–5, May 2010.

[216] J. R. Bradford, Y. Hey, T. Yates, Y. Li, S. D. Pepper, and C. J. Miller, "A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling," *BMC genomics*, vol. 11, p. 282, May 2010.

[217] S. Liu, L. Lin, P. Jiang, D. Wang, and Y. Xing, "A comparison of rna-seq and high-density exon array for detecting differential gene expression between closely related species," *Nucleic acids research*, Sep 2010.

[218] A. Agarwal, D. Koppstein, J. Rozowsky, A. Sboner, L. Habegger, L. W. Hillier, R. Sasidharan, V. Reinke, R. H. Waterston, and M. Gerstein, "Comparison and calibration of transcriptome data from rna-seq and tiling arrays," *BMC Genomics*, vol. 11, p. 383, Jan 2010.

[219] K. L. Stone, R. D. Bjornson, G. G. Blasko, C. Bruce, R. Cofrancesco, N. J. Carriero, C. M. Colangelo, J. K. Crawford, J. M. Crawford, N. C. daSilva, J. D. Deluca, J. I. Elliott, M. M. Elliott, P. J. Flory, E. J. Folta-Stogniew, E. Gulcicek, Y. Kong, T. T. Lam, J. Y. Lee, A. Lin, M. B. LoPresti, S. M. Mane, W. J. McMurray, I. R. Tikhonova, S. Westman, N. A. Williams, T. L. Wu, Z. Hongyu, and K. R. Williams, "Keck foundation biotechnology resource laboratory, yale university," *Yale J Biol Med*, vol. 80, pp. 195–211, Dec 2007.

[220] F. Hsu, W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, and D. Haussler, "The ucsc known genes," *Bioinformatics*, vol. 22, pp. 1036–46, May 2006.

[221] R. A. Miller and N. L. Nadon, "Principles of animal use for gerontological research," *J Gerontol A Biol Sci Med Sci*, vol. 55, pp. B117–23, Mar 2000.

[222] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev, "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas," *Nat Biotechnol*, vol. 28, pp. 503–10, May 2010.

[223] Z. Wen, C. Wang, Q. Shi, Y. Huang, Z. Su, H. Hong, W. Tong, and L. Shi, "Evaluation of gene expression data generated from expired affymetrix genechip® microarrays using maqc reference rna samples," *BMC Bioinformatics*, vol. 11 Suppl 6, p. S10, Jan 2010.

[224] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nat Rev Cancer*, vol. 8, pp. 37–49, Jan 2008.

[225] G. L. G. Miklos and R. Maleszka, "Microarray reality checks in the context of a complex disease," *Nat Biotechnol*, vol. 22, pp. 615–21, May 2004.

[226] J. P. A. Ioannidis, "Microarrays and molecular research: noise discovery?," *Lancet*, vol. 365, pp. 454–5, Jan 2005.

[227] J. P. A. Ioannidis, "Why most published research findings are false," *Plos Med*, vol. 2, p. e124, Aug 2005.

[228] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics*, vol. 21, pp. 171–8, Jan 2005.

[229] L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T.-M. Chu, F. M. Goodsaid, L. Pusztai, J. D. Shaughnessy, A. Oberthuer, R. S. Thomas, R. S. Paules, M. Fielden, B. Barlogie, W. Chen, P. Du, M. Fischer, C. Furlanello, B. D. Gallas, X. Ge, D. B. Megherbi, W. F. Symmans, M. D. Wang, J. Zhang, H. Bitter, B. Brors, P. R. Bushel, M. Bylesjo, M. Chen, J. Cheng, J. Cheng, J. Chou, T. S. Davison, M. Delorenzi, Y. Deng, V. Devanarayan, D. J. Dix, J. Dopazo, K. C. Dorff, F. Elloumi, J. Fan, S. Fan, X. Fan, H. Fang, N. Gonzaludo, K. R. Hess, H. Hong, J. Huan, R. A. Irizarry, R. Judson, D. Juraeva, S. Lababidi, C. G. Lambert, L. Li, Y. Li, Z. Li, S. M. Lin, G. Liu, E. K. Lobenhofer, J. Luo, W. Luo, M. N. McCall, Y. Nikolsky, G. A. Pennello, R. G. Perkins, R. Philip, V. Popovici, N. D. Price, F. Qian, A. Scherer, T. Shi, W. Shi, J. Sung, D. Thierry-Mieg, J. Thierry-Mieg, V. Thodima, J. Trygg, L. Vishnuvajjala, S. J. Wang, J. Wu, Y. Wu, Q. Xie, W. A. Yousef, L. Zhang, X. Zhang, S. Zhong, Y. Zhou, S. Zhu, D. Arasappan, W. Bao, A. B. Lucas, F. Berthold, R. J. Brennan, A. Buness, J. G. Catalano, C. Chang, R. Chen, Y. Cheng, J. Cui, W. Czika, F. Demichelis, X. Deng, D. Dosymbekov, R. Eils, Y. Feng, J. Fostel, S. Fulmer-Smentek, J. C. Fuscoe, L. Gatto, W. Ge, D. R. Goldstein, L. Guo, D. N. Halbert, J. Han, S. C. Harris, C. Hatzis, D. Herman, J. Huang, R. V. Jensen, R. Jiang, C. D. Johnson, G. Jurman, Y. Kahlert, S. A. Khuder, M. Kohl, J. Li, L. Li, M. Li, Q.-Z. Li, S. Li, Z. Li, J. Liu, Y. Liu, Z. Liu, L. Meng, M. Madera, F. Martinez-Murillo, I. Medina, J. Meehan, K. Miclaus, R. A. Moffitt, D. Montaner, P. Mukherjee, G. J. Mulligan, P. Neville, T. Nikolskaya, B. Ning, G. P. Page, J. Parker, R. M. Parry, X. Peng, R. L. Peterson, J. H. Phan, B. Quanz, Y. Ren, S. Riccadonna, A. H. Roter, F. W. Samuelson, M. M. Schumacher, J. D. Shambaugh, Q. Shi, R. Shippy, S. Si, A. Smalter, C. Sotiriou, M. Soukup, F. Staedtler, G. Steiner, T. H. Stokes, Q. Sun, P.-Y. Tan, R. Tang, Z. Tezak, B. Thorn, M. Tsyganova, Y. Turpaz, S. C. Vega, R. Visintainer, J. von Frese, C. Wang, E. Wang, J. Wang, W. Wang, F. Westermann, J. C. Willey, M. Woods, S. Wu, N. Xiao, J. Xu, L. Xu, L. Yang, X. Zeng, J. Zhang, L. Zhang, M. Zhang, C. Zhao, R. K. Puri, U. Scherf, W. Tong, and R. D. Wolfinger, "The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models," *Nat Biotechnol*, vol. 28, pp. 827–38, Aug 2010.

[230] K. A. Baggerly, K. R. Coombes, and E. S. Neeley, "Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer," *J Clin Oncol*, vol. 26, pp. 1186–7; author reply 1187–8, Mar 2008.

[231] J. P. A. Ioannidis, "Is molecular profiling ready for use in clinical decision making?," *The Oncologist*, vol. 12, pp. 301–11, Mar 2007.

[232] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc Natl Acad Sci USA*, vol. 103, pp. 5923–8, Apr 2006.

Publications

R.R. Kitchen, V.S. Sabine, A.A. Simen, J.M. Dixon, J.M.S. Bartlett, A.H. Sims. (2011) Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments. BMC Genomics (submitted)

V. Ogilvie, M. Passmore, L. Hyndman, L. Jones, B. Stevenson, A. Wilson, H. Davidson, **R.R. Kitchen**, R.D. Gray, P. Shah, E.W. Alton, J.C. Davies, D.J. Porteous, A.C. Boyd. (2011) Differential global gene expression in cystic fibrosis nasal and bronchial epithelium. Genomics (in press)

K.A. Bordner, **R.R. Kitchen**, B.C. Carlyle, E.D. George, M.C. Mahajan, S.M. Mane, J.R. Taylor, A.A. Simen. (2011) Parallel declines in cognition, motivation, and locomotion in aging mice: association with immune gene upregulation in the medial prefrontal cortex. Experimental Gerontology (e-pub ahead of print)

B. Kelmendi, M. Holsbach-Beltrame, A.M. McIntosh, L. Hilt, E.D. George, **R.R. Kitchen**, B.C. Carlyle, C. Pittenger, V. Coric, S. Nolen-Hoeksema, G. Sanacora, A.A. Simen (2011) Association of polymorphisms in HCN4 with mood disorders and obsessive compulsive disorder. Neuroscience Letters. 496(3):195-9

K.A. Bordner, E.D. George, B.C. Carlyle, A. Duque, **R.R. Kitchen**, T.T. Lam, C.M. Colangelo, K.L. Stone, T.B. Abbott, S.M. Mane, A.C. Nairn, A.A. Simen. (2011) Functional genomic and proteomic analysis reveals disruption of myelin-related genes and translation in a mouse model of early life neglect. Frontiers in Psychiatry. 2:18

R.R. Kitchen, M. Kubista, A. Tichopad. (2010) Statistical aspects of quantitative real-time PCR experiment design. Methods. 50(4):231-6

R.R. Kitchen, V.S. Sabine, A.H. Sims, E.J. Macaskill, L. Renshaw, J.S. Thomas, J.I. van Hemert, J.M. Dixon, J.M.S. Bartlett. (2010) Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles. BMC Genomics. 11(1):134

A. Tichopad, T. Bar, L. Pecen, **R.R. Kitchen**, M. Kubista, M.W. Pfaffl. (2010) Quality control for quantitative PCR based on amplification compatibility test. Methods. 50(4):308-12

A. Tichopad, **R. Kitchen**, I. Riedmaier, C. Becker, A. Stahlberg, M. Kubista. (2009) Design and optimization of reverse-transcription quantitative PCR experiments. Clinical Chemistry. 255(10):1816-23